Putting the Human Back in the AutoML Loop

Iordanis Xanthopoulos*
University of Crete
Greece
jordan.xanthopoulos@gmail.com

Ioannis Tsamardinos†‡
University of Crete
Gnosis DA
Greece
tsamard.it@gmail.com

Vassilis Christophides§
University of Crete
Greece
christop@csd.uoc.gr

Eric Simon
SAP France
France
eric.simon@sap.com

Alejandro Salinger
SAP SE
Germany
alejandro.salinger@sap.com

ABSTRACT

Automated Machine Learning (AutoML) is a rapidly rising sub-field of Machine Learning. AutoML aims to fully automate the machine learning process end-to-end, democratizing Machine Learning to non-experts and drastically increasing the productivity of expert analysts. So far, most comparisons of AutoML systems focus on quantitative criteria such as predictive performance and execution time. In this paper, we examine AutoML services for predictive modeling tasks from a user's perspective, going beyond predictive performance. We present a wide palette of criteria and dimensions on which to evaluate and compare these services as a user. This qualitative comparative methodology is applied on seven AutoML systems, namely Auger.AI, BigML, H2O's Driverless AI, Darwin, Just Add Data Bio, RapidMiner, and Watson. The comparison indicates the strengths and weaknesses of each service, the needs that it covers, the segment of users that is most appropriate for, and the possibilities for improvements.

KEYWORDS

AutoML, machine learning services, qualitative evaluation

1 INTRODUCTION

Automated Machine Learning (AutoML) is becoming a separate, independent sub-field of Machine Learning, that is rapidly rising in attention, importance, and number of applications [23, 35]. AutoML goals are to completely automate the application of machine learning, statistical modeling, data mining, pattern recognition, and all advanced data analytics techniques. As an end result, AutoML could potentially democratize ML to non-experts (Citizen Data Scientists), boost the productivity of experts, shield against statistical methodological errors, and even surpass manual expert analysis performance (e.g., by using meta-level learning [11]). Finally, AutoML could improve replicability of analyses, sharing of results, and facilitate collaborative analyses.

To clarify the term AutoML, we consider the minimal requirements to be the ability to return (a) a predictive model that can be applied to new data, and (b) an estimate of predictive performance of that model, given a data source, e.g., a 2-dimensional matrix (*tabular data*). Thus, do-it-yourself tools that allow you to graphically construct the analysis pipeline (e.g. Microsoft's Azure ML [31]) are excluded. In addition, we distinguish between *libraries* and *services*. The former require coding and typically offer just the minimal requirements, namely return a model and a performance estimation. AutoML services, on the other hand, include a user interface and strive to democratize ML not only to coders, but to anybody with a computer; they typically offer a much wider range of functionalities.

Algorithmically, AutoML encompasses techniques regarding hyper-parameter optimization (HPO, [3, 48]), algorithm selection (CASH, [22]), automatic synthesis of analysis pipelines [36], performance estimation [53], and meta-level learning [54], to name a few. In addition, an AutoML system could not only automate the modeling process, but also the steps that come before and after. Pre-analysis steps include data integration, data preprocessing, data cleaning, and data engineering (feature construction). Post-analysis steps include interpretation, explanation, and visualization of the analysis process and the output model, model production, model monitoring, and model updating. The ideal AutoML system should only require the human to specify the data source(s), their semantics, and the goal of the analysis to create and maintain a model into production indefinitely.

Given the importance and potential of AutoML, several academic and commercial libraries, as well as services have appeared. The first AutoML system was the academic Gene Expression Model Selector (GEMS) [46]. Recent works formulate the AutoML problem [56, 57], introduce techniques and frameworks for creating new AutoML tools [6, 45], survey the existing ones [43, 57] and comparatively evaluate them [42, 51, 55]. This is a a technically challenging task requiring the availability of a plethora of datasets with different characteristics [14], extensive computational time, ability to set time-limits to all software and many others (see [18] for a discussion on the set up and results of the AutoML Challenge Series).

AutoML strives to take the human expert out of the ML loop; but, unfortunately, it seems the majority of AutoML surveys and evaluations also take the human user out of the loop, focusing solely on predictive performance and ignoring the user experience for the most part. Still, some exceptions can be found. In

*This work was done while the author was working at SAP SE
†Ioannis Tsamardinos is CEO of Gnosis Data Analysis, which created JAD Bio (JAD)
‡The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 617393
§Work of the author was supported by the Institute of Advanced Studies of the University Cergy-Pontoise under the Paris Seine Initiative for Excellence ("Investissements d'Avenir" ANR-16-IDEX-0008).

[26], an interactive environment is proposed, emphasizing on user-centric aspects of AutoML. Moreover, in [49] a brief qualitative evaluation on AutoML services and libraries is presented, mainly regarding their ML capabilities.

The contribution of this paper is to provide a user-centric framework for comparing AutoML services. We define a set of qualitative criteria, spanning across six categories (Estimates, Scope, Productivity, Interpretability, Customizability, and Connectivity) that highlight user-experience beyond predictive performance when selecting or evaluating AutoML services. Using this framework we evaluated seven such services, namely Auger.AI [2], BigML [4], H2O's Driverless AI [19], Darwin [7], Just Add Data Bio [1], RapidMiner [39], and Watson [24]. The comparison is meant to indicate the strengths, weaknesses, scope, and usability of the services, indicating the needs it covers, the tasks it is most appropriate for, and the opportunities for improvement. To the best of our knowledge no other survey or benchmarking paper proposed the aforementioned qualitative criteria and methodology for evaluating AutoML services and libraries.

2 AUTOML SERVICES CONSIDERED

In the present evaluation study we consider seven current AutoML service platforms that offer a free trial version, so we could base it on first-hand experience. All of these services, specialize on tabular data, helping us apply the qualitative criteria on all of them. **It was conducted from 01/12/2019 until 07/12/2019 and we used the live versions of the services at the time.** In alphabetical order, the services are:

- **Auger.AI[2]**: A new service, going live in 2019, Auger.AI boasts to have high accuracy and a well-implemented API to help users run experiments with ease.
- **BigML [4]**: One of the oldest ML services, BigML supports AutoML tasks and offers extended support, a custom programming language and a cloud infrastructure for the user.
- **Darwin [7]**: SparkCognition's new AutoML service, providing the users with convenient tools to speed-up their ML tasks.
- **Driverless AI (DAI) [19]**: One of the most well-known AutoML services, DAI supports various ML tasks and also has advanced interpretability mechanisms.
- **Just Add Data Bio (JAD) [1]**: JAD was launched in November 2019 focusing on the analysis of molecular biological data (small-sample, high-dimensional) with emphasis on feature selection.
- **RapidMiner Studio (RM) [39]**: The oldest AutoML service used in our evaluation, RM provides multiple tools to its users and supports user-created components. We are looking into the standard version, not including the available user-created add-ons.
- **IBM's Watson (Watson) [24]**: Watson contains multiple components, but here we focus on the *AutoAI experiment* toolkit[1], being closer to what we define as AutoML service for tabular data.

Due to registration fees, we were not able to include in our benchmark recent services such as Google AutoML Tables[2]. Regarding Data Robot[3], we were not able to obtain the free trial licence advertised on their website.

[1] https://www.ibm.com/cloud/watson-studio/autoai
[2] https://cloud.google.com/automl-tables/

3 QUALITATIVE CRITERIA

To qualitatively evaluate the seven AutoML services, we present *32* user-centric qualitative criteria spanning across six different categories. The criteria are partitioned in the following categories. The *Estimates* category is concerned with metrics and estimates' properties about the predictive power of the final model. The *Scope* criteria describe the applicability scope of a service mainly in terms of data types and ML predictive tasks. The *Productivity* category is concerned with the ease of use, while *Interpretability* is concerned with the ability to interpret the results of the analysis. The last two categories are *Customizability* of the analysis and *Connectivity* of the service. The criteria are graded on a 4-level scale. F(ail) (✗), *C* for fulfilling the basic requirements of the criterion, *B* for providing additional functionalities and *A* for achieving a level that should satisfy most users in our opinion.

3.1 Estimates

Criteria for *Estimates* (Table 1), concern the wealth and depth of estimated quantities regarding the predictive model. *ROC curves* are a useful visualization for interpreting the performance of a classification model and are widely used by the ML community. We grade with *B* the services that output ROC curves (Auger.AI, BigML and RM) and with *A* the ones which also output performance metrics for different points on the curve (DAI, JAD and Watson). In addition to the out-of-sample estimate of predictive performance, a service should be able to report the uncertainty of this estimation (criterion *STD/CI calculation* in Table 1 standing for standard deviation and confidence interval respectively). With *B*, we grade the services that only calculate the STD (BigML, DAI and RM) and with *A* the ones calculating the whole probability distribution of performance and its confidence intervals, a richer piece of information (JAD). Regarding *Label Predictions* on new data, the services that support either individual samples predictions or batch predictions are graded with *B* (Darwin), and the ones supporting both with *A* (the rest of the services). For binary classification tasks, the services able to generate *Label probability estimations* get an *A* (all services except Auger.AI). Overall, JAD has a full score on all the criteria, followed by DAI and RM.

3.2 Scope

Scope criteria (Table 1) cover the range of input data that can be analyzed. When it comes to *Outcome types*, services able to handle binary (classification), multi-class (classification), continuous (regression) and censored time-to-event outcomes (survival analysis) score *A* (JAD), while the ones not handling survival analyses score *B* (the rest of the services). Regarding *Predictor types*, the services which support all the standard tabular data and also text or time-series data are graded with *A* (all services except for JAD), while the ones only supporting the former with *B* (JAD). The term *Clustered data* (not to be confused with clustering of data) in statistics refers to samples that are naturally grouped in clusters (or groups) of samples that may be correlated given the predictors. Examples include matched case-control data in medicine and repeated measurements taken on the same subject or client. With *A*, we grade the services able to handle clustered data (DAI and JAD). It is important to mention the absence of

[3] https://www.datarobot.com/

Table 1: Estimates and Scope criteria.

	Criteria	Auger.AI	BigML	DAI	Darwin	JAD	RM	Watson
Estimates	ROC curves	B	B	A	✗	A	B	A
	STD/CI calculation	✗	B	B	✗	A	B	✗
	Label predictions	A	A	A	B	A	A	A
	Label probability estimations	✗	A	A	A	A	A	A
Scope	Outcome types	B	B	B	B	A	B	B
	Predictor types	A	A	A	A	B	A	A
	Clustered data handling	✗	✗	A	✗	A	✗	✗
	Missing values handling	A	A	A	A	A	A	A

clustered data and repeated measurements handling from most of the services. Essentially, most services assume independently and identically distributed (i.i.d.) data reducing their scope. Finally, we grade a service's ability to handle missing data with *A* (all services). In this category, DAI and JAD lead with the highest score.

3.3 Productivity

The *Productivity* criteria (Table 2) concern the ease of use and boost of user productivity. We start off with *Data manipulation* functionalities available to prepare and manipulate the input data before analysis. Grade *B* goes to the services providing the user with custom data partitioning and preprocessing recommendations (DAI and Darwin) and grade *A* to the services that additionally provide data merging, filtering and sub-sampling (BigML, JAD, RM, Watson). About *Pipeline automation*, the services where the best model is automatically selected according to pre-specified user preferences (e.g., maximize AUC) score *A* (DAI, Darwin, JAD and Watson). The services producing a ranking of all tried models instead and require the user to select the one that satisfies their criteria the best score *B* (Auger.AI, BigML and RM). On one hand, ranking all the models arguably provides richer information to the user, on the other, it does reduce automation and could confuse the non-expert. So, our grading in this criterion is admittedly subjective. We next grade the ability to *Early stop or pause* an analysis. The services able to do both score *A* (RM) and in case they have implemented either one but not the other, they score *B* (the rest of the services). When it comes to *Collaboration features*, we grade a service with *A* if it has implemented mechanisms to create custom organizations and teams to allow sharing of resources, such as data and analyses (all services except DAI and Darwin). Lastly, about *Documentation and support*, the services providing e-mail support score *C* (JAD). If they also deliver extensive documentation to the user, they score *B* (Auger.AI and Darwin) and when they additionally have direct technical support and user forums, their score is *A* (BigML, DAI, RM and Watson). In general, Productivity is a category emphasized by all services, making it relatively straightforward to any user to complete an ML analysis.

3.4 Interpretability

Interpretability criteria (Table 2) is arguably on the most important categories for selecting an AutoML service[32]. The criteria concern (a) Exploring and visualizing the data (*Data visualization*) before conducting the analysis. (b) Monitoring the execution of the analysis progress (*Progress report*). (c) Understanding and interpreting how the final model functions (*Final model interpretation*). A particular means to understanding of results is through *Feature selection*, which deserves its own criterion, along with the available mechanisms for the *Final feature set interpretation*. (d) Understanding and validating the process that took place during the analysis (*Analysis exploration*). Regarding *Data visualizations* prior to the analysis, a service which only provides histograms, scores *C* (JAD). If it also implements correlation plots and data heatmaps, its score is *B* (BigML). The services with more options get *A* (DAI, RM and Watson). During the analysis (*Progress report*), if a service only reports the completion percentage, it gets the grade *C* (Darwin). When it shows additionally a performance estimation of the best model and keeps track of the analysis procedure, its grade is *B* (BigML and JAD). The highest grade (*A*) goes to the services that also show variable importance rankings, generated models ranking and hardware usage (Auger.AI, DAI, RM and Watson).

Once the analysis is complete, the AutoML service should be able to explain how the final model works. This adds transparency to the model and pinpoints possible flaws or bias in its decision making, making it more trustworthy. The interpretability of the results is a subdomain of ML with increasing popularity and every year multiple new mechanisms are introduced [9, 33]. We have selected a set of such mechanisms and grade the AutoML services based on how many of them they have implemented. The mechanisms are: a) the confusion matrix, which is created based on the predictions made during the training phase, to help the user understand what type of errors are produced by the final model; b) report of the performance of the final model using multiple performance metrics; c) residuals visualization, i.e. the difference between observed and predicted values of the data; d) PCA procedure [44] to highlight strong patterns of the data and visualize them on a 2-D space; e) visualization of the final model, when this is possible; f) techniques to explain the predictions in case of a complex final model (e.g. LIME-SUP [21], K-LIME, a variant of LIME [40], decision tree surrogate models [8], etc.). When the service has implemented at least 2 of the above mechanisms, its corresponding grade is *C* (Darwin and Watson), while for a service with more than 2 available mechanisms, its grade is *B* (Auger.AI, BigML, RM). The grade *A* is reserved for the services with more than 4 of the aforementioned mechanisms implemented (DAI and JAD).

Feature selection is often the *primary* goal of an analysis. It leads to simpler models that require fewer measurements to provide a prediction, which may be important in several applications. Most importantly however, *feature selection is used as a tool for knowledge discovery* [28] to gain intuition and insight into the

Table 2: Productivity and Interpretability criteria. ✜: only for certain models

	Criteria	Auger.AI	BigML	DAI	Darwin	JAD	RM	Watson
Productivity	Data manipulation	✗	A	B	B	A	A	A
	Pipeline automation	B	B	A	A	A	B	A
	Early stop or pause	B	✗	B	B	B	A	B
	Collaboration features	A	A	✗	✗	A	A	A
	Documentation and support	B	A	A	B	C	A	A
Interpretability	Data visualization	✗	B	A	✗	C	A	A
	Progress report	A	B	A	C	B	A	A
	Final model interpretation	B	B	A	C	A	B	C
	Feature selection	✗	C	C	✗	A	B	✗
	Final feature set interpretation	C	B	A	C	A	B	C
	Analysis exploration	A✜	B	B	✗	B	A	A

problem (hence, its inclusion in the interpretability category). A pharmacologist is not only interested in predicting cancer metastasis but also in the molecules involved in the prediction to identify drug targets; a business person is interested in the quantities that affect customer attrition to devise new promotions and advertisements. Such reasoning is theoretically supported by the fact that feature selection has been connected to the causal mechanisms that generate the data [50]. It is defined as the problem of identifying a *minimal-size* feature subset that *jointly* (multivariately) leads to an *optimal* prediction model (see [17] for a formal definition). Thus, feature selection removes not only irrelevant, but also redundant features. In some data distributions, there may be multiple solutions to the feature selection. For example, due to low sample size the truly best feature subset may be statistically indistinguishable from slightly sub-optimal feature subsets. Or, it could be the case there is informational redundancy that leads to feature subsets that are equally predictive. While all solutions are equivalent in terms of predictive performance, *returning all solutions is important when feature selection is used as a tool for knowledge discovery*.

The services which offer single feature selection functionality, score *C* (BigML and DAI). BigML treats feature selection as a preprocessing step, before the modeling process and the estimation of performance protocol. This approach is methodologically wrong and leads to overestimating performance (see [20], page 245). There are different notions of multiple feature selection. When a service returns several feature subsets as options, but does not provide any theoretical guarantees of statistical equivalence, its grade is *B* (RM). On the other hand, when a service returns several feature subsets that lead to models with statistically indistinguishable performance from the optimal, its grade is *A* (JAD). Feature selection by itself is not enough. The services should also provide users with mechanisms for interpreting and understanding how each feature in the final set affects and contributes to the decision making of the final model. We base our grading on a set of *Final feature set interpretation* mechanisms and how many of them each AutoML service has implemented. The mechanisms are: a) random forest feature importance ranking of the participating features [5]; b) LOCO feature importance [27]; c) partial dependence plots (PDPs) [12]; d) SHAP plots [29]; e) ICE plots [15]; f) a report of the standardized individual and cumulative importance of the participating features; g) the actual standardized coefficient for each feature, in the case of a linear final model; h) information about the resulted feature sets, in the case of multiple feature selection. A service that has implemented at least 1 of these mechanisms, is graded with *C* (Auger.AI, Darwin and Watson). If more than 2 mechanisms are available, the service's grade is *B* (BigML, RM) and the grade *A* is reserved for the services with 4 or more mechanisms (DAI, JAD).

Expert analysts would often like to verify the correctness and completeness of the analysis that took place. It is not only the results (model) that should not be treated as a black-box, but also how these results were obtained. A service which displays an *Analysis exploration* graph, to help the users understand the methods used in each step scores *A* (Auger.AI, RM and Watson). If the service displays all pipelines that were tried, in the form of list instead of as a graph, its score is *B* (BigML, DAI and JAD). When it comes to analysis interpretation, DAI and JAD seem to be the best choice, providing the user with advanced mechanisms for understanding the final results. Some services, do not provide any information about which analysis pipelines they tried; the analysis process is essentially a black box to the user. We note that in our opinion, there is room for improvement regarding interpretability for most of the services.

3.5 Customizability

The *Customizability* category (Table 3) grades the ability of the services to customize analysis according to user choices and preferences. About *Time budget*, we grade with *B* the services giving the ability to impose a non-strict time limit on an analysis (Auger.AI, BigML and JAD) and with *A* the ones which allow setting a strict time limit (DAI and Darwin). Our take on this subject is that every service should give the ability to pose a strict time budget, as an analysis can be part of a bigger project, running under specific time restrictions. Moving to the hardware *Resources budget*, if a service allows the user to select a preset hardware configuration, it scores *B* (Watson) and if it allows setting up the exact hardware specifications, *A* (DAI and JAD). Next, we consider the *Customization of analysis components*, i.e. the ability to choose the methods and algorithms to try, along with their hyperparameters, in each step of the ML pipeline. If the user is able to fully customize the included components, the service gets *A* (Auger.AI, BigML, DAI and RM). If the service provides the user with a set of limited settings, it gets *B* (Darwin, JAD and Watson).

A service that allows the user to *Enforce final model interpretability*, is graded with *B* (JAD) and if it provides additional interpretability settings, with *A* (DAI). Another customization

Table 3: Customizability and Connectivity criteria. ✧: for RM server, not RM studio

	Criteria	Auger.AI	BigML	DAI	Darwin	JAD	RM	Watson
Customizability	Time budget	B	B	A	A	B	✗	✗
	Resources budget	✗	✗	A	✗	A	✗	B
	Analysis components customization	A	A	A	B	B	A	B
	Enforce Model Interpretability	✗	✗	A	✗	B	✗	✗
	Feature selection options	✗	A	A	✗	A	B	✗
	Visualizations customization	✗	A	B	✗	✗	A	A
Connectivity	Service deployment	✗	A	A	✗	✗	A✧	✗
	3rd party storage connection	A	A	A	✗	✗	A	A
	API access	A	A	A	A	A	A	A
	Downloadable results	A	A	A	✗	B	A	B
	Analysis components contribution	B	A	A	✗	✗	A	B
	Model deployment	A	A	A	A	✗	A	A
	Visualizations exportability	✗	B	B	✗	B	A	A

criterion is about the available *Feature selection options*. If the AutoML service allows the user to select the exact number of selected features, it is graded with *A* (BigML, DAI and JAD) and if it allows the user to set certain parameters, such as the effort put in feature selection, with *B* (RM). Finally, we also consider the *Visualizations customization* options. When a service gives the user the ability to set user-specific thresholds on certain visualizations, its grade is *B* (DAI). If the user can fully customize the resulted visualizations (e.g. changing the axes, titles, legend, colors), the service's grade is *A* (BigML, RM and Watson). In general, when it comes to customizability, DAI has a clear edge over the competition, giving the users options to fine-tune and setup an analysis according to their needs. We distinguish two different schools of thought on this category. On one hand, services such as DAI, let the user fully customize the algorithms and hyperparameter values to search during an analysis. On the other hand, services like JAD provide the user with a few preference choices that do not require expert knowledge of ML. The first approach empowers an expert analyst but it may be intimidating to the non-expert user. There is a fine line between providing enough choices to an expert to fully customize an analysis and achieve better results and providing too many choices that make the process complex and easy to break. For this reason, we would recommend to equip AutoML services with some kind of warning system that can actually detect when the selected setup might create problems and notify the user accordingly.

3.6 Connectivity

The *Connectivity* criteria (Table 3) grade the options offered to connect a service with external tools and resources. First, regarding the *Service's deployment* at an external infrastructure, the services supporting it score *A* (BigML, DAI and RM). The ones able to *Connect to 3rd party storage providers* also get an *A* (all except from Darwin and JAD). Furthermore, all services have implemented their own *API* (grade *A*). We also look into the *Downloadable results* options. In the case where only part of the results are downloadable, the services are graded with *B* (JAD and Watson) while the ones allowing the user to download all the results and also generate a summary report, with (A) (all services except JAD and Watson). A user might be interested in *Adding custom components* to the AutoML service. If it is allowed to the user to add components through a service's API, the service is graded with *B* (Auger.AI and Watson). If the service has moreover implemented a complete system for user-defined components, by creating their own marketplace or extensions library, its grade is *A* (BigML, DAI and RM). Creating the best final model does not always suffice, as the user will probably want to deploy it in an external service and use it for new data predictions. Most of the participating services, have added various model deployment options (grade *A*) (all except JAD). The currently implemented ideas are to use data transfer libraries, e.g. cURL (Auger.AI, Watson), create actionable models (BigML, Darwin, RM) or scoring pipelines (DAI). All of the above provide the same functionality; predicting labels on new unseen data. Finally, when writing reports or papers with the results, the visualizations need to be exported. The services which provide less than 3 export options score *B* (BigML, DAI and JAD) and those with more, score *A* (RM and Watson). Taking a look at the participating services, most of them cover the majority of the proposed criteria. The export formats available for data visualizations are static in all tools, an area that could greatly be improved. Additionally, we find the lack of connections to public repositories, such as OpenML [52] important, as they can be useful to a user who is interested in conducting ML analyses for academic reasons.

4 LIMITATIONS AND DISCUSSION

Admittedly, the current study has several limitations. We take the opportunity to discuss some in depth, pointing to important open issues and future work. First of all, we were not able to evaluate every known AutoML service.

Estimates: While all services provide estimated quantities from the data, the major question remains: **are the estimates returned correct and reliable**? Statistical estimations are particularly challenging with low samples; even more so with high dimensional data. Is performance overestimated, standard deviations underestimated, probabilities of individual predictions uncalibrated, feature importance's accurate, or multiple feature subsets returned not statistically equivalent? *Which AutoML services return reliable results one can trust, and which ones are actually misleading the user and potentially harmful*? In case of medical applications, overestimating performance or confidence in a prediction (uncalibrated predicted probabilities) is dangerous and could impact human health, while in business applications it may have significant monetary costs. Such questions require

significant experimentation with all services to answer. Experimentation should be performed on datasets with a wide range of characteristics, e.g., sample size, number of features, percentage of missing values, mixture of types of predictors (continuous, discrete, ordinal, zero-inflated, etc.), outcomes, etc. to provide a full quantitative picture of the pros and cons of each service and its correctness properties. Unfortunately, most quantitative evaluations are currently performed on datasets with a limited range of such characteristics or are restricted by time limitations.

Scope: In this paper, we are only concerned with predictive modeling (supervised learning) tasks and not other ML categories. Each different task would require a separate set of criteria that applies to it. *We do note, however, that BigML, DAI, RM, and Watson also support clustering, anomaly detection, and some NLP tasks* which are useful to numerous users. A major limitation of our scope grading is that it misses important criteria concerning the maximum volume of data a service can handle in reasonable time or memory resources, both in terms of number of features, samples, or their combination (total volume). Unfortunately, we are not able to test the limits of each service as we are confined to analyses that run on the free trial versions. However, regarding the scalability with respect to feature size, we note that almost all services have difficulty scaling to thousands of features. JAD on the other hand, was created to scale up to the feature size of typical multi-omics datasets that can reach up to hundreds of thousands of features.

Productivity/Interpretability: Although, we presented a first qualitative assessment, a true measure of productivity increase requires an extensive user study with representative datasets spanning a wide-range of characteristics (in terms of the number of features and samples). In such a user-study, one should measure how much productivity has improved over manual scripting, eventually by trading off learning performance, and how much insight has been gained by the interpretation tools offered by each service. To assess how an AutoML tool performs against human experts Kaggle[4] and other ML competitions could be exploited. As data and tasks are specific for a competition problem, solutions by human experts usually take the top positions as they apply domain-specific knowledge and sometimes create custom methods and mechanisms to help them win these competitions. Still, AutoML tools that have been tested on such tasks, achieve comparable performance. AutoML tools are becoming more and more sophisticated, by automating an increasing number of tasks in ML pipelines (e.g., feature engineering), while supporting meta-level learning techniques. This can lead to minimizing the gap between human experts and AutoML in competitive environments [45] and aid in producing high quality ML models for both commercial and academic purposes.

There are several other criteria categories that are missing from the present methodology, due to space limitations. These include *model monitoring and maintenance* that regards functionalities to maintain a model into production [30], such as monitor the health of the production model, raise alarms when there is a drift in the data distribution, automatically re-train and update the model, and others. As ML systems move from computer-science laboratories into the open world, their *accountability* [13] and *auditing* [10] becomes a high priority problem. In this respect, we need a deep understanding of the ML system behavior and its failures. Current evaluation methods such as single-score error metrics and confusion matrices provide aggregate views of system performance that hide important shortcomings. Understanding details about failures is important for finding ways for improvement, communicating the reliability of systems in different settings and for specifying appropriate human oversight and engagement [34].

Finally, we would like to mention that each category could be expanded with many more criteria. Only the criteria that were addressed by at least one of the services were included. Functionalities that were not addressed by any of the services examined are missing. One example is the ability to handle continuous signals and streaming data [38].

[4]https://kaggle.com

5 CONCLUSION

AutoML has made tremendous progress since its first embodiment in the GEMS system. Several AutoML services are already available, routinely analyzing business and scientific data for thousands of users. They do increase productivity and allow non-experts to perform sophisticated ML analyses. Our prediction is that within a few years, most of data analysis will involve the use of an AutoML service or library; scripting as a means to manual ML analysis will gradually become obsolete or pass to the next level, where it is customizing and invoking AutoML functionalities.

The proposed criteria intend to turn the spotlight back onto the human user. Users do not only consider learning performance when choosing a service. They also consider a plethora of other criteria such as the ones presented. One of the most important ones is interpretability of results. Users are rarely satisfied with just a predictive model; they also seek to understand the patterns in their data. Thus, results should not be a black-box, but explained, visualized, and interpreted. Users need to examine the analysis process and ensure its correctness or optimality: AutoML should automate, not obfuscate. The analysis process should be transparent, verifiable, and customizable by the user. Some of the AutoML services examined, clearly abide to these principles but some fail in this set of criteria. Arguably, it is perhaps interpretation of results and ease-of-use that will determine the success of an AutoML service, and not necessarily predictive performance.

Current AutoML systems mostly focus on tabular, iid-sampled data. Obviously however, most of the world's data is not in this format or sampled as iid. Ultimately, AutoML competes with the human expert not only in learning performance but in scope and the range of problems it can handle. There are ongoing efforts to develop AutoML solutions for regression or anomaly detection tasks in time-series, time-course data, and streaming data (e.g., Microsoft Azure [31], Yahoo EGADS [25], Facebook Prophet [47]), or to generate features from relational tables or CSV/JSON files [16]. Future AutoML systems should also automate more data preparation tasks including data cleaning (e.g. error correction and deduplication) [41] and support ML tasks such as reinforcement, transfer and federated learning, or causal modeling [37] to name a few. Still, interpreting the results of the analysis in each category is quite challenging and probably requires a different, specialized set of methods. There is a long road ahead, where ML is entering a new generation of systems and algorithms, but an exciting road indeed.

REFERENCES

[1] Gnosis Data Analysis. 2019. *Just Add Data Bio.* https://www.jadbio.com/.
[2] Auger.AI. 2019. *Auger.AI.* https://auger.ai/.

[3] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2546–2554. http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf

[4] BigML. 2012. *BigML*. https://bigml.com/.

[5] Leo Breiman. 2017. *Classification and regression trees*. Routledge.

[6] Yi-Wei Chen, Qingquan Song, and Xia Hu. 2019. Techniques for Automated Machine Learning. *CoRR* abs/1907.08908 (2019). arXiv:1907.08908 http://arxiv.org/abs/1907.08908

[7] Spark Cognition. 2019. *Darwin*. https://www.sparkcognition.com/product/darwin/.

[8] Mark W. Craven and Jude W. Shavlik. 1995. Extracting Tree-Structured Representations of Trained Networks. In *Proceedings of the 8th International Conference on Neural Information Processing Systems (NIPS'95)*. MIT Press, Cambridge, MA, USA, 24–30.

[9] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[10] Amitai Etzioni and Oren Etzioni. 2016. Designing AI Systems That Obey Our Laws and Values. *Commun. ACM* 59, 9 (Aug. 2016), 29–31.

[11] Matthias Feurer and Frank Hutter. 2018. Towards Further Automation in AutoML. In *ICML 2018 AutoML Workshop*.

[12] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[13] Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. 2019. Explainable AI in Industry. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 3203–3204.

[14] Pieter Gijsbers, Erin LeDell, Janek Thomas, Sébastien Poirier, Bernd Bischl, and Joaquin Vanschoren. 2019. An Open Source AutoML Benchmark. *CoRR* abs/1907.00909 (2019). arXiv:1907.00909 http://arxiv.org/abs/1907.00909

[15] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24, 1 (2015), 44–65.

[16] Google. 2019. *AutoML Tables*. https://cloud.google.com/automl-tables/.

[17] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

[18] Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, Alexander Statnikov, Wei-Wei Tu, and Evelyne Viegas. 2019. *Analysis of the AutoML Challenge Series 2015–2018*. Springer International Publishing, Cham, 177–219. https://doi.org/10.1007/978-3-030-05318-5_10

[19] H2O. 2017. *Driverless AI*. https://www.h2o.ai/products/h2o-driverless-ai/.

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[21] Linwei Hu, Jie Chen, Vijayan Nair, and Agus Sudjianto. 2018. Locally Interpretable Models and Effects based on Supervised Partitioning (LIME-SUP). (06 2018).

[22] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. 2011. Sequential Model-Based Optimization for General Algorithm Configuration. In *Proceedings of the conference on Learning and Intelligent OptimizatioN (LION 5)*. 507–523.

[23] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (Eds.). 2018. *Automated Machine Learning: Methods, Systems, Challenges*. Springer. In press, available at http://automl.org/book.

[24] IBM. 2015. *IBM Watson Studio*. https://www.ibm.com/watson.

[25] Nikolay Laptev, Saeed Amizadeh, and Ian Flint. 2015. Generic and Scalable Framework for Automated Time-Series Anomaly Detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1939–1947. https://doi.org/10.1145/2783258.2788611

[26] Doris Jung-Lin Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya Parameswaran. 2019. A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *Data Engineering* (2019), 58.

[27] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1094–1111.

[28] Huan Liu and Hiroshi Motoda. 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, USA.

[29] Scott Lundberg, Gabriel Erion, and Su-In Lee. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. (02 2018).

[30] Jorge G Madrid, Hugo Jair Escalante, Eduardo F Morales, Wei-Wei Tu, Yang Yu, Lisheng Sun-Hosoya, Isabelle Guyon, and Michèle Sebag. 2018. Towards AutoML in the presence of Drift: first results. In *Workshop AutoML 2018 @ ICML/IJCAI-ECAI*. Pavel Brazdil, Christophe Giraud-Carrier, and Isabelle Guyon, Stockholm, Sweden. https://hal.inria.fr/hal-01966962

[31] Microsoft. 2015. *Azure Machine Learning Studio*. https://studio.azureml.net/.

[32] Christoph Molnar. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

[33] Christoph Molnar. 2019. *Interpretable machine learning*. Lulu. com.

[34] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018*. 126–135.

[35] Meghana Padmanabhan, Pengyu Yuan, Govind Chada, and Hien Van Nguyen. 2019. Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction. In *Journal of clinical medicine*.

[36] Magnus Palmblad, Anna-Lena Lamprecht, Jon Ison, and Veit Schwämmle. 2018. Automated workflow composition in mass spectrometry-based proteomics. *Bioinformatics* 35, 4 (2018), 656–664.

[37] J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

[38] Fábio Pinto, Marco O. P. Sampaio, and Pedro Bizarro. 2019. Automatic Model Monitoring for Data Streams. *CoRR* abs/1908.04240 (2019). http://arxiv.org/abs/1908.04240

[39] RapidMiner. 2006. *RapidMiner*. https://rapidminer.com/.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[41] Vraj Shah and Arun Kumar. 2019. The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning (DEEM'19)*. Association for Computing Machinery, New York, NY, USA, Article Article 11, 4 pages. https://doi.org/10.1145/3329486.3329499

[42] Zeyuan Shang, Emanuel Zgraggen, Benedetto Buratti, Ferdinand Kossmann, Philipp Eichmann, Yeounoh Chung, Carsten Binnig, Eli Upfal, and Tim Kraska. 2019. Democratizing Data Science Through Interactive Curation of ML Pipelines. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19)*. ACM, New York, NY, USA, 1171–1188. https://doi.org/10.1145/3299869.3319863

[43] Radwa El Shawi, Mohamed Maher, and Sherif Sakr. 2019. Automated Machine Learning: State-of-The-Art and Open Challenges. *CoRR* abs/1906.02287 (2019). arXiv:1906.02287 http://arxiv.org/abs/1906.02287

[44] Jonathon Shlens. 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100* (2014).

[45] Micah J. Smith, Carles Sala, James Max Kanter, and Kalyan Veeramachaneni. 2019. The Machine Learning Bazaar: Harnessing the ML Ecosystem for Effective System Development. *CoRR* abs/1905.08942 (2019). arXiv:1905.08942 http://arxiv.org/abs/1905.08942

[46] Alexander Statnikov, Ioannis Tsamardinos, Yerbolat Dosbayev, and Constantin F Aliferis. 2005. GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *International journal of medical informatics* 74, 7-8 (2005), 491–503.

[47] Sean Taylor and Benjamin Letham. 2017. Forecasting at Scale. *The American Statistician* 72 (09 2017). https://doi.org/10.1080/00031305.2017.1380080

[48] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2012. Auto-WEKA: Automated selection and hyper-parameter optimization of classification algorithms. *CoRR, abs/1208.3719* (2012).

[49] Anh Truong, Austin Walters, Jeremy Goodsitt, Keegan Hines, Bayan Bruss, and Reza Farivar. 2019. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. (2019). http://arxiv.org/abs/1908.05557 cite arxiv:1908.05557.

[50] Ioannis Tsamardinos and Constantin F Aliferis. 2003. Towards principled feature selection: relevancy, filters and wrappers.. In *AISTATS*.

[51] Lukas Tuggener, Mohammadreza Amirian, Katharina Rombach, Stefan Lörwald, Anastasia Varlet, Christian Westermann, and Thilo Stadelmann. 2019. Automated Machine Learning in Practice: State of the Art and Recent Results. *CoRR* abs/1907.08392 (2019). arXiv:1907.08392 http://arxiv.org/abs/1907.08392

[52] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations* 15, 2 (2013), 49–60. https://doi.org/10.1145/2641190.2641198

[53] Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics* 7, 1 (2006), 91.

[54] Ricardo Vilalta and Youssef Drissi. 2002. A Perspective View and Survey of Meta-Learning. *Artificial Intelligence Review* 18, 2 (01 Jun 2002), 77–95. https://doi.org/10.1023/A:1019956318069

[55] Ziqiao Weng. 2019. From Conventional Machine Learning to AutoML. *Journal of Physics: Conference Series* 1207 (apr 2019), 012015. https://doi.org/10.1088/1742-6596/1207/1/012015

[56] Quanming Yao, Mengshuo Wang, Hugo Jair Escalante, Isabelle Guyon, Yi-Qi Hu, Yu-Feng Li, Wei-Wei Tu, Qiang Yang, and Yang Yu. 2018. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. *CoRR* abs/1810.13306 (2018). arXiv:1810.13306 http://arxiv.org/abs/1810.13306

[57] Marc-André Zöller and Marco F. Huber. 2019. Survey on Automated Machine Learning. (2019). arXiv:1904.12054 http://arxiv.org/abs/1904.12054