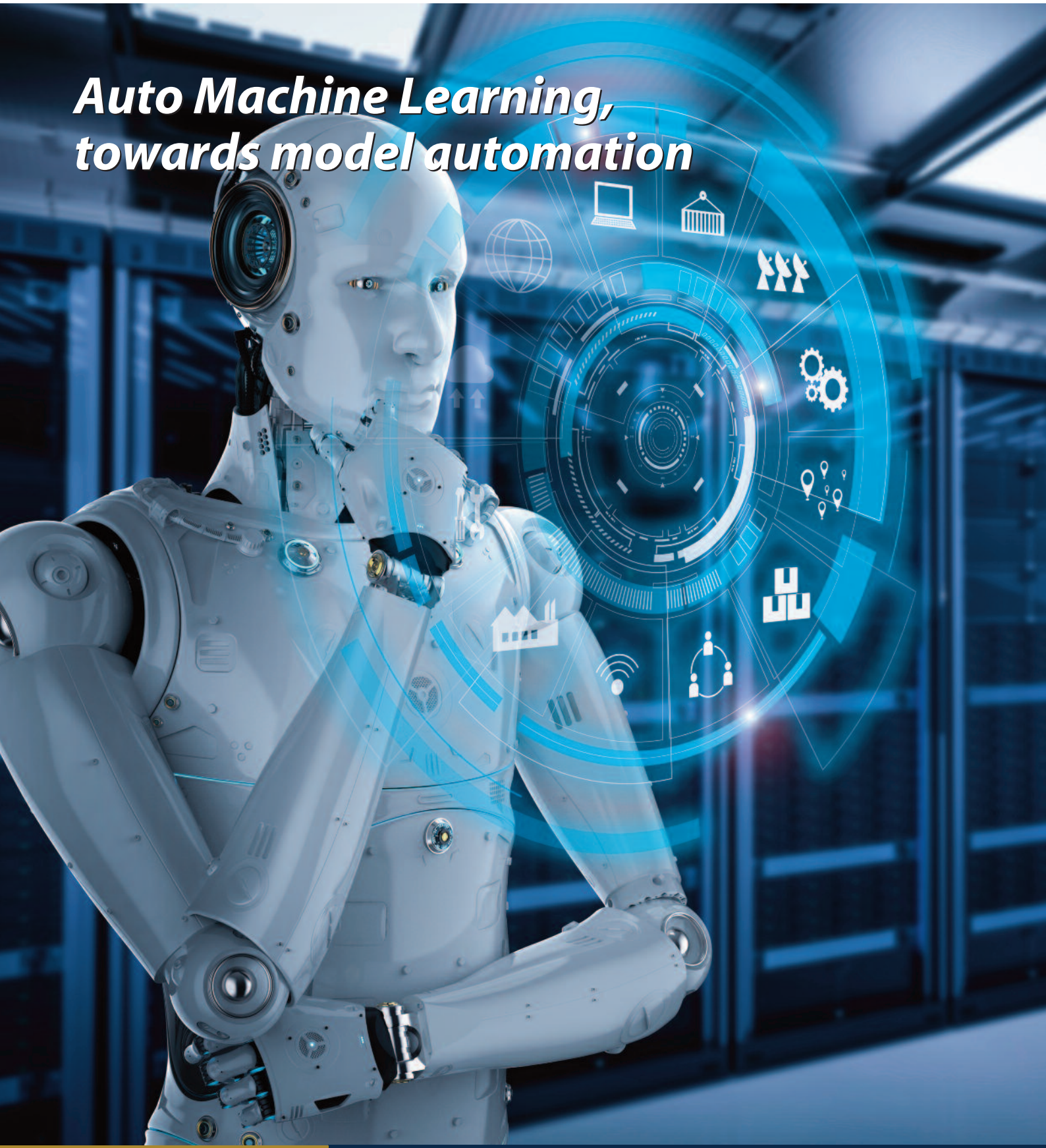
ManagementSolutions
Making things happen

Auto Machine Learning, towards model automation



R&D

www.managementsolutions.com

Design and Layout
Marketing and Communication Department
Management Solutions

Photographs
Photographic archive of Management Solutions
iStock

Contents

Introduction

"A half-dozen monkeys provided with typewriters would, in a few eternities, produce all he books in the British Museum"

– Jorge Luis Borges[1]

A mathematical model is, in a way, a simplification of reality that takes advantage of the information available to systematize decision-making. This simplification allows hypotheses on the behavior of both variables and systems to be evaluated through their summary representation under a set of postulates, usually based on data and applying inference criteria. Its main purpose is to explain, analyze or predict the behavior of a variable.

The current revolution in modeling techniques, coupled with increased computing power and more accessible and greater data storage capacity, has radically changed the way models have been built in recent years. This revolution has been a key factor that has stimulated the use of these new techniques not only in decision-making processes where traditional approaches were used, but also in areas where the use of models was not so common. Finally, in some industries, such as the financial sector, the use of models has also been driven by regulation. Standards such as IFRS 9 and 13 or Basel II have promoted the use of internal models with the aim of adding sensitivity and making the calculation of accounting impairment or financial risks more sophisticated.

Although it may appear otherwise, the most common modeling techniques currently used in the business field do not have a recent origin. Specifically, linear and logistic regressions date from the 19th century. However, for some time now there have been significant developments in new algorithms that, while aimed at enhancing how patterns are found in the data, also introduce new challenges such as the need for improved interpretability techniques. The use of these new mathematical models in computing is a scientific discipline known as machine learning, since it allows systems to learn and find patterns without being explicitly programmed to do so.

There are multiple definitions of machine learning, two of the most illustrative being those of Arthur Samuel and Tom Mitchell. For Arthur Samuel[2] , machine learning is "the field of study that gives computers the ability to learn without being explicitly programmed", while for Tom Mitchell[3] it is "a program that learns from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E". These two definitions are usually related to unsupervised and supervised learning, respectively[4].

As a consequence, the appetite to properly understand and draw conclusions from data has increased dramatically. However, at the same time, implementing these methods has required changing multiple aspects in organizations[5] and is, in turn, a source of potential risk due to development or implementation errors or inappropriate use

Advanced modeling improves business and operational processes, or even facilitates the emergence of new business models. An example can be found in the financial sector, where new digitization methodologies are modifying the current value proposition, but also adding new services. According to a survey conducted by the Bank of England and the Financial Conduct Authority on almost 300 companies in the financial and insurance sectors, two thirds of the participants used machine learning in their processes[6]. Machine learning techniques are frequently used in typical control tasks, such as money laundering prevention (AML) or fraud detection, the analysis of cybersecurity-related threats, and in business processes such as customer classification, recommendation systems or customer service through the use of chatbots. They are also used in credit risk management, pricing, operations and insurance underwriting.

Other sectors have seen a similar level of development. The use of machine learning models is common in industries such as manufacturing, transportation, medicine, justice or the retail and consumer goods sectors. This has caused investment in

[1] Jorge Luis Borges, "La biblioteca total" (1939).). Argentine writer, poet, essayist and translator whose works include Ficciones and El Aleph.
[2] Samuel, 1959.
[3] Mitchell, 1997.
[4] Management Solutions, 2018.
[5] Ibid.
[6] Bank of England, 2019.

companies dedicated to artificial intelligence to increase from a total of $ 1.3 billion in 2010 to $ 40.4 billion in 2018[7] (see Figure 1). The expected return justifies this investment: 63% of companies that have adopted Machine Learning models have reported increased revenues, with approximately half of them reporting an increase exceeding 6%. Likewise, 44% of companies reported cost savings, with approximately half of them achieving savings over 10%.[8].

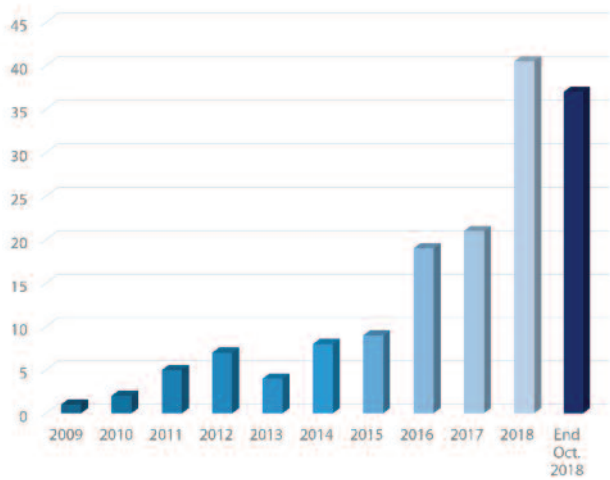
Of the different changes made by organizations to adapt to this new paradigm, talent recruitment and retention is still central. To begin with, companies have needed to enlarge their teams specializing in machine learning. The demand for professionals in this field increased by 728% between 2010 and 2019 in the United States[10] (see Figure 2), with a qualitative change in the demand for general data scientist skills and knowledge also being notable.

But this demand for generic machine learning and data scientist skills is not always the case: in order to analyze the increasing amounts of data available using increasingly sophisticated tools, requirements for skills have become more specific to include knowledge of different programming languages such as Python, R, Scala or Ruby, the ability to deal with databases in big data architectures, knowledge of cloud computing, advanced mathematical and statistical knowledge, and specialized postgraduate training. Consequently, many different jobs in the market have become hard to fill due to their highly specific skill set requirements.. In addition, the rate at which companies are generating data means that, even with a stable supply of data scientists, the current recruitment solution is not scalable.

However, it is not only necessary to have specialist teams, but also to implement new development procedures, review validation methods, revise and assess models in the validation and audit areas, and make important cultural changes in other areas for this implementation to be effective. The use of these new processes creates a chain reaction that affects the entire model life cycle, including notably the identification,

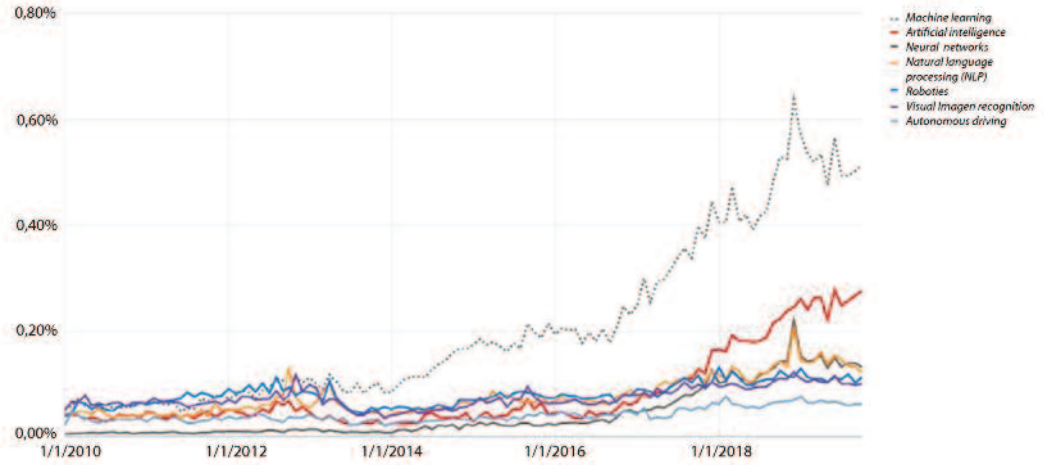
[7] For investments above 400, 000 US dollars. Stanford University, 2019.
[8] Statista, 2019.
[9] Stanford University, 2019.
[10] Ibid.

Figure 1: annual investment in AI companies (in billion dollars)[9].



Source: Stanford University, 2019.

Figure 2: increase in the demand for machine learning and artificial intelligence skill sets.



Source: Ibid.

management and governance of model risk[11]. Many of these models also require supervisory approval, as in the financial industry (e.g. capital or provision models), or in the pharmaceutical industry, which adds new challenges such as the need to ensure the interpretability of the models used, as well as to develop other elements of model confidence.

Another notable aspect of the investment in machine learning methods is that its development is uneven across organizations: the need to undergo validation, audit and approval processes established by regulations, or the requirement to maintain specific documentation standards, is creating differences in the implementation of internal models across firms. According to the EBA's big data and analytics report[12], financial institutions are adopting digital transformation programs or promoting the use of machine learning techniques in areas such as risk mitigation (including automated scoring, operational risk management or fraud) and Know Your Customer processes. However, "although the use of machine learning may represent an opportunity to optimize capital, from a prudential framework perspective it is premature to consider the use of machine learning techniques appropriate for determining capital requirements"[13].

[11] Management Solutions, 2014.
[12] European Banking Authority, 2020.
[13] Ibid.

There are also operational risks that are difficult to detect, such as those arising from human error during model implementation, or those related to data storage security, which should be properly managed to ensure machine learning systems are used in a suitable environment. An example of this is the framework established by the European Commission in these cases, covering different aspects of the modeling process[14]. Finally, and also due to both regulatory and management considerations, models need to operate reliably and be used ethically so that they can be trusted in the decision-making process. The EBA's proposal in this regard, based on seven pillars of trust[15], is of particular interest: ethics, interpretability, avoidance of bias, traceability, data protection and quality, security and consumer protection. These issues have also been identified as key elements by universities and business spheres[16].

In this context, different model development tasks demand very different times: the tasks prior to and complementary to analysis also require a large amount of time and resources to prepare, clean and generally process the data; 60% of a data scientist's time is spent cleaning data and organizing information, while 9% and 4% of their time is spent on knowledge discovery tasks and algorithm tuning, respectively[17]. All this drives the need to change the way in which model development, validation and implementation is approached, to take advantage of the new techniques while solving the difficulties associated with their use, as well as mitigating any potential risks.

For the reasons outlined above, there is a clear trend towards automating processes related to the use of advanced analytics techniques – generally called automated machine learning or AutoML, whose aim is not only to automate those tasks where heuristic processes are limited and easily automatable, but also to allow for more automated, ordered and traceable algorithm and pattern search processes to be generated. According to Gartner[18], more than 50% of data science tasks will be automated by 2025.

Furthermore, this trend towards automation offers a number of opportunities, such as the ones offered by the automation systems architecture in terms of workflow design model inventory, or component validation. Automated machine learning systems integrate various tools to develop models, also reducing cost, development time and system implementation errors.

AutoML systems and methods seek, among other things, to:

- reduce the time spent by data scientists on developing models through the use of machine learning techniques, and even to allow non-data scientist teams to develop machine learning algorithms;
- improve model performance, as well as model traceability

[14] European Comission, 2020.
[15] European Banking Authority, 2020.
[16] For example, the iDanae Chair, which resulted from a collaboration between the Polytechnic University of Madrid and Management Solutions, has published newsletters on interpretability (iDanae Chair, 3T-2019) and ethics in artificial intelligence (iDanae Chair, 4T-2019).
[17] According to a survey carried out by the CrowdFlower Artificial Intelligence platform (CrowdFlower, 2017).
[18] Gartner, 2019.

and comparability against manual hyperparameter search techniques;

- challenge models developed using other approaches; iv) leverage the investment made in terms of both time and resources to develop codes and improve the system's components efficiently and with greater traceability; and

- simplify the validation of models and facilitate their planning.

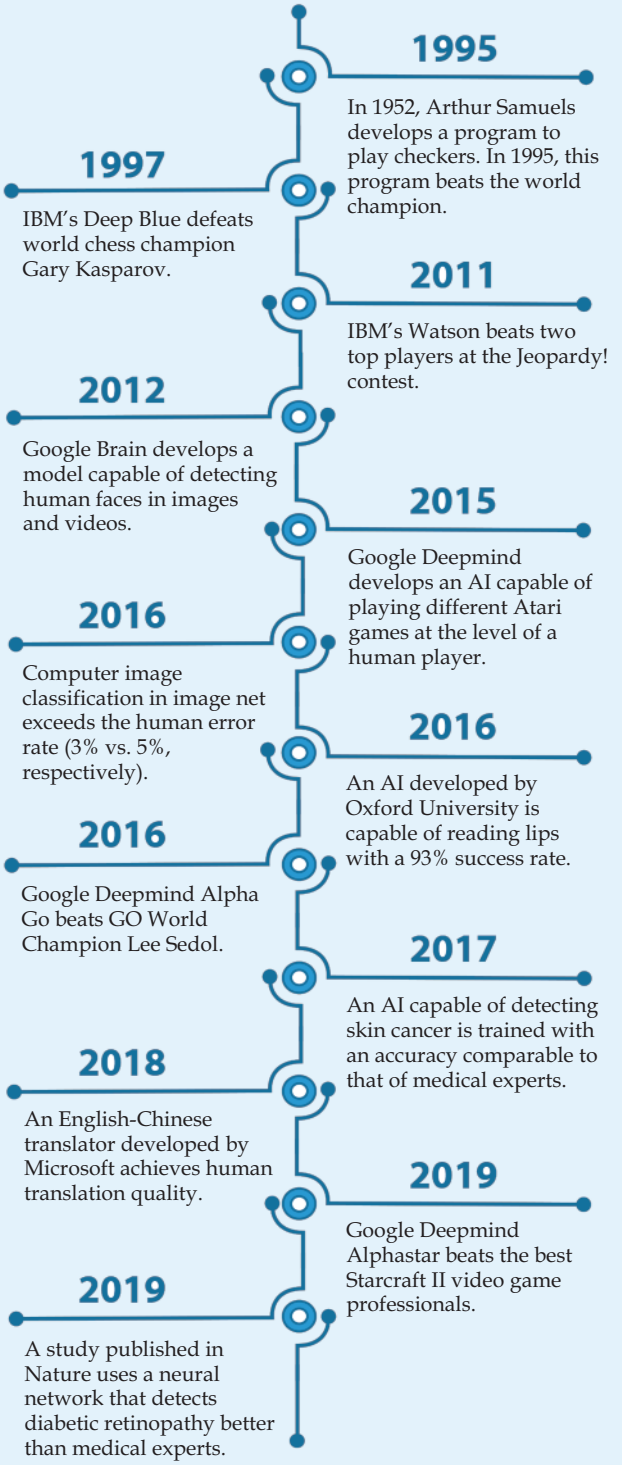
Against this backdrop, this document aims to describe the key elements of AutoML systems. For this purpose, it has been structured in three sections, with three objectives:

- In the first section, the factors explaining the move towards the automation of machine learning processes are analyzed, as are the reasons underlying the development of AutoML systems, through both their componentization and automation.

- The second section provides a descriptive view of the main AutoML frameworks, and explains what approaches are being followed, both in the academic field and in practical experiences aimed at automating modeling processes through machine learning techniques.

- Finally, the third section aims to illustrate the results of AutoML system development, presenting as a case study a competition organized by Management Solutions in early 2020. The aim of this competition, aimed at MS professionals, was to design an Automated Machine Learning model.

Key milestones in ML development



EBA's report on big data and advanced analytics

The European Banking Authority (EBA) has published a report on big data and advanced analytics with the aim of publicizing their use in the European financial sector, as well as sharing the bank's insights on (i) the four key challenges identified in the development, implementation and adoption of big data advanced analytics, (ii) the key elements of trust on which a big data and advanced analytics framework should be based, and (iii) key trends, opportunities and risks arising from the use of these solutions.

I. Key pillars of a big data and advanced analytics framework

II. Elements of trust

III. Key trends, opportunities and risks in the use of big data and advanced analytics

Key pillars of a big data and advanced analytics framework

DATA MANAGEMENT

- Data types and data sources: structured, unstructured and semi-structured data.
- Data security and data protection: protect the confidentiality, integrity and availability of data.
- Data quality: ensure data quality throughout the whole Big data and Advanced Analytics (BD&AA) lifecycle.

ANALYTICS METHODOLOGY

- Data preparation: feature engineering and feature importance analysis
- Analytics: model training, tuning, validation and selection, testing and deployment.
- Operations: model monitoring and update.

TECHNOLOGICAL INFRASTRUCTURE

- Infrastructure: networking resources to transmit Big Data
- Data platform: manages all the data used by an advanced analytics system
- Processing: provides the necessary software to support the implementation of advanced analytics applications

ORGANIZATION AND GOVERNANCE

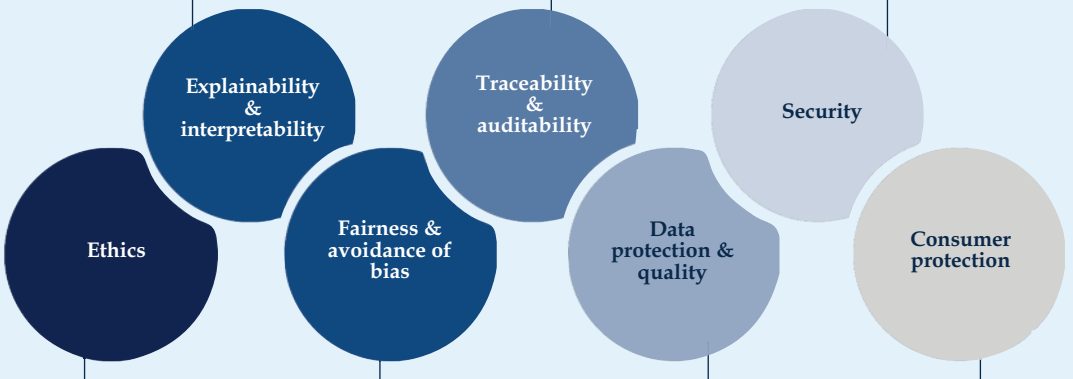
- Internal governance structures and measures: governance structure, strategy and risk management, transparency, external development and outsourcing.
- Skills and knowledge: level of understanding of management team and senior management, second and third line, developers of advanced analytics-enabled systems.

Elements of trust

A key consideration to building trustworthy models is that it should be possible to explain and interpret them, and they should be transparent, properly understood and provide clear justifications.

Using traceable solutions assists in tracking all steps, criteria and choices in the process. This makes it possible to track all processes resulting in the decisions made by the model, which in turn ensures the system is auditable.

It is important to maintain a technical watch on the latest security attacks and related defence techniques and ensure that governance, oversight and the technical infrastructure are in place for effective Information and Communications Technology (ICT) management.



Developing, deploying and using artificial intelligence solutions requires adhering to some fundamental ethical principles.

Fairness requires that the model ensure the protection of groups against discrimination. To ensure fairness, the modelshould be free from bias.

Data should be adequately protected with a trustworthy BD&AA system and its quality needs to be taken into account throughout the BD&AA lifecycle.

A trustworthy BD&AA system should respect consumers' rights and protect their interests.

Key trends, opportunities and risks in the use of big data and advanced analytics

Key observations

- Institutions are at different stages of BD&AA development. Typical use cases found in fraud detection are CRM and process automation.
- More reliance on internal data, rather than external data or social media. Incorporation of open source solutions. Limited use of complex algorithms.
- Different level of integration and governance of advanced analytics in organizations.
- Increasing reliance on technology companies for the provision of both infrastructure and cloud services.

Key opportunities

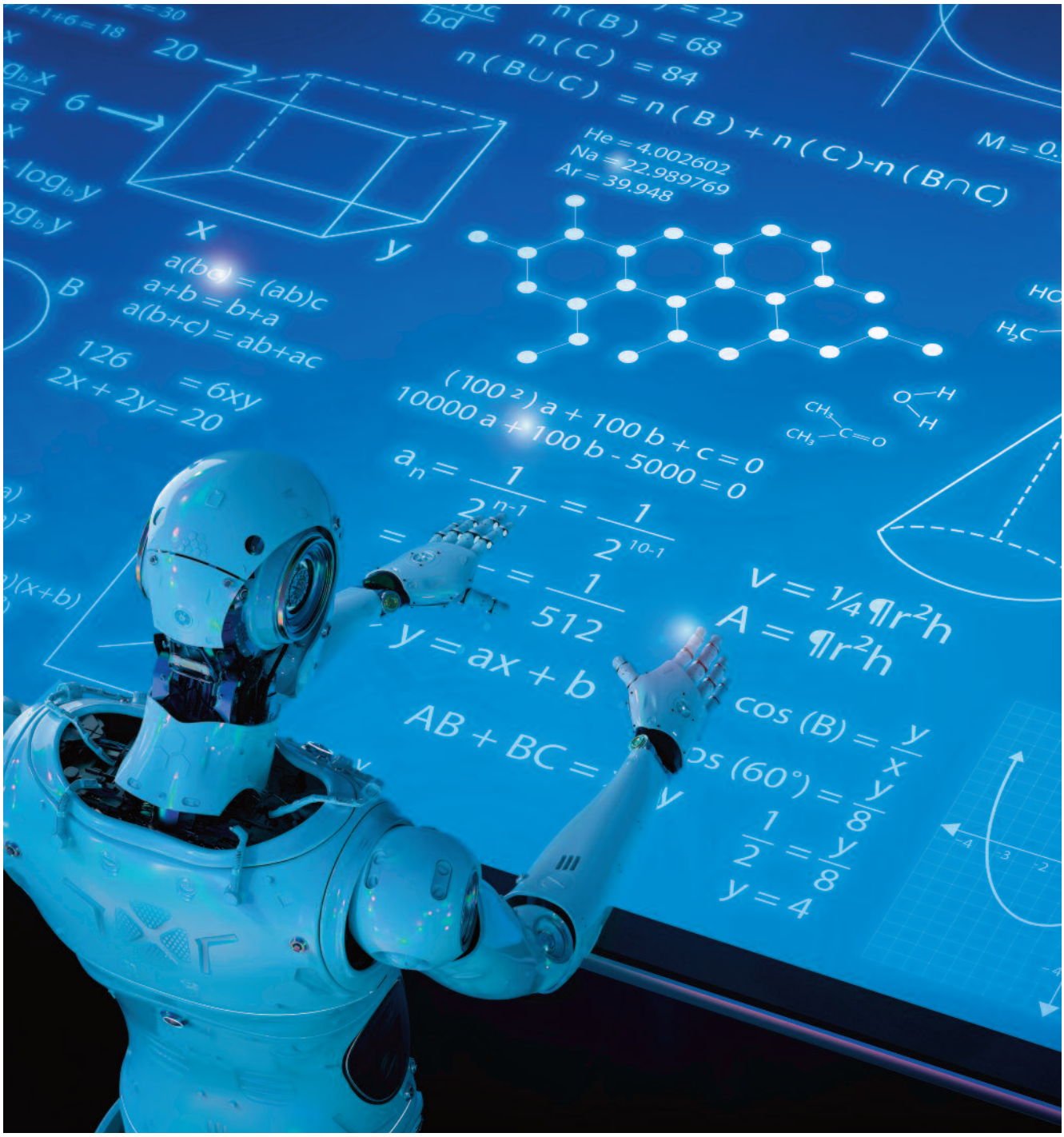
- Financial services customers from leisure and retail sectors expect a more personalized service. There is trust in the financial sector regarding GDPR compliance.
- Improvement of customer satisfaction and use of insights to improve the offering, reduce churn, optimize processes, and assist risk mitigation and fraud detection.
- Many possible uses and opportunities arising from the use of interpretable models.

Key risks and proposed guidance

- Model output can be complex, non deterministic, and correct according to a probability measure, which could harm both organizations and customers. It has to be ensured that the outputs of these systems do not violate the ethical standards of organizations. In addition, there should be human-in-the-loop involvement in decisions, therefore there is a need for proper employee training.
- The implementation of a governance and methodological framework on BD&AA could promote its responsible use, which should include appropriate documentation, sufficient justification, and other explainability and traceability techniques, including the use of traceable solutions. The explainability should be based on a risk-based approach.
- There is a need for model accuracy and regular monitoring.
- The use of ML solutions could give rise to ICT risks: data security, model security, data quality, change management, and business continuity and resilience.
- As a consequence of the reliance on open source frameworks, or on tools and systems developed by third parties, both their potential risks (lack of third-party knowledge and control, vendor lock-in, concentration risk, model maintenance, etc.) and the liability, which will always remain with the entity, must be assessed.
- Finally, the importance of data quality, protection and security is emphasized, both for regulatory purposes (including compliance with GDPR) and to ensure the adequacy of the model.

Executive summary

"Strictly speaking, one inmortal monkey would suffice "
Jorge Luis Borges



The context of machine learning model automation

1. The incorporation of big data and advanced analytics techniques into the economy is changing the way information is used. By combining different data and business analysis knowledge, analysis capabilities have increased radically, although at the same time there are potential risks in connection with their development or implementation, their inappropriate use or overconfidence in their application.

2. To be able to take advantage of the potential of these new techniques, firms are transforming their ways of working. These changes directly affect model development and validation, but also other processes such as those related to IT structures, the selection, training and retention of specialist profiles and, in a broader sense, the work culture.

3. There are operational risks involved that are difficult to detect and, in some cases, regulations on data use, data quality, and data related processes, as well as on the use and interpretability of models, that determine the need for traceable processes and for recurrent validations and model reviews.

4. In this context, there is a clear trend towards the automation of processes related to the application of advanced analytics techniques, the aim of which is not only to automate tasks where heuristic processes are limited and easily automatable, but also to allow more automated, ordered and traceable model development.

5. All this leads to a reduction in the time spent by data scientists on supplementary and repetitive tasks, as well as to increased access to these techniques by non-specialist teams, greater model performance, traceability and comparability, reuse of code developments in specific projects, improved techniques, and even improved validation processes, including the development of challenger models.

Towards model automation

6. There are multiple workflow automation challenges in the modeling process. Some of these challenges arise from the need to have processes in place that ensure proper data loading, and others are related to model development, validation, implementation and monitoring. A further challenge is to ensure data building process traceability and interpretability, as well as adequate governance to allow for this process to be embedded in the business as usual, and to comply with existing regulations.

7. Regarding data processing, 60% of a data scientist's time is spent cleaning and organizing the information, and there is a long way to go before these processes are fully automated.

8. As for the modeling workflow, there are two ways in which this process can be automated, and they are usually combined: the componentization of the different processes into segregated elements, and automatic execution of these components, systematizing them through pre-established rules and statistical techniques.

9. Componentization is based on separating modeling tasks into different parts, and conducting their programming and development independently. Each of these components receives a specific input and executes a specific task.

10. The advantages of componentization are process standardization, increased quality and efficiency, specialized development, improved usability, and greater scalability.

11. On the other hand, automation of the model building process is based on the use of automatic criteria for selecting the model's attributes, so that the procedure can be replicated and audited. It also ensures that the final selection has been made through a process that ensures optimal predictive power given constraints.

12. The advantages of search automation are optimized hyperparameter selection process, the generalization of modeling problems, parameter search spaces that are adapted to each problem, and the possibility of experimenting outside the usual ranges.

Machine learning process automation frameworks

13. In practice, these processes are automated by (i) systematizing most of the aspects related to the analysis and prior treatment of the data, including the transformation of variables and their preselection, (ii) creating a search space for possible models and parameters, as well as a model development and selection process that avoids both overfitting[19] and underfitting[20], and (iii) automating the implementation of interpretability techniques.

14. There are many options for putting these systems into production, which fall under either model-based schemes or data-driven approaches.

15. Model building and evaluation processes are fundamentally based on two components: the optimizer and the evaluator.

16. The optimizer produces and updates parameter combinations within the limit of possibilities defined depending on the model and the data used. The evaluator then measures the performance of the options proposed by the optimizer, and may influence the search strategy based on the results.

17. The optimizer uses different techniques to find the best configuration. These techniques can be simple (grid search, random search, evolutionary algorithms, Bayesian optimization) or experience-based (meta-learning, transfer learning).

18. The evaluator checks whether the configuration given by the optimizer is optimal. There are different optimization approaches, such as (i) early stopping, where the evaluator stops evaluating if performance is very low in the first iterations, (ii) reuse, based on using configurations utilized in previous training, or (iii) using surrogate models in the evaluation.

19. In this type of system, the challenge is to include prior knowledge, such as business knowledge or expert criteria, as well as to develop systems that cover the entire model-building process.

20. An alternative to using a system based on the interaction of an optimizer and an evaluator is to search for neural architectures (NAS). This technique, used in language modeling or image classification, simultaneously performs the three tasks required for automation: it determines the search space and the search strategy for this space, and estimates model performance.

21. Although performance is high under the above approach, it is more difficult to explain why some configurations are reached and whether these configurations can be extended for use in other types of problems.

22. Despite there being much room for improvement, AutoML systems have currently reached a stage of development that can compete and beat human experts in machine learning, having become an essential tool that allows data scientists to focus on analysis tasks taking place before and after model development itself.

23. The biggest challenges, both for hyperparameter optimization systems and for NAS methods, are related to interpretability, reproducibility, and reusing configurations from previous exercises for better user interaction.

[19] Refers to a model that has been adjusted too closely to the training data set, and therefore does not achieve satisfactory results on other data sets.

[20] Refers to a model that has not been sufficiently adjusted to the training data set, and therefore does not achieve satisfactory results on other data sets.

AutoML competitions: an AutoML approach exploration tool

24. To deepen the understanding and implementation of AutoML approaches, competitions have been organized to compare methodologies. An example of these competitions is offered in Chapter 5.

25. In the case of the AutoML competition ran by Management Solutions, participants used approaches similar to those discussed above, e.g. grid, random search, genetic algorithms or Bayesian searches, to produce models.

26. Some useful conclusions were drawn from that exercise: i) data processing was quite homogeneous, showing that a number of standardized techniques are used in the industry; ii) dimensionality reduction was used by most participants to significantly bring down the computational cost; iii) AutoML systems improve by optimizing the entire pipeline, using stacking models, or parallelizing tasks across multiple cores.

Final insights

27. Currently the configuration of machine learning models depends significantly on a priori and manual adjustments, which can lead to suboptimal results due to both overfitting and underfitting, depending on the size of the dataset and the techniques used. Model overfitting is still common with some techniques, which suggests there is room for improvement in AutoML systems in some specific cases.

28. Although AutoML approaches have developed to high levels, there are still shortcomings related to the fact that the pipeline is not fully automated, to the absence or lack of objectivity in some of the decisions, and to the search space.

29. Another challenge is to allow non-expert profiles access to AutoML environments for them to directly interact with these methods and systems so that business intuition can be incorporated, or so that they may directly evaluate the interpretability of models. Finally, interpretability remains one of the open questions in AutoML systems.

30. As for recent advances, these are more common in feature engineering optimization and in model selection, to the detriment of data processing or preparation.

31. Last, it is expected that AutoML systems will emerge as a fundamental tool that can modify the work carried out before and after model development, on producing AutoML components and systems, and on solving specific problems where an AutoML system does not achieve good results

Towards model automation

"And if you play it for a hundred years, or a thousand years or a hundred thousand, the law of chances tells us that a poem will probably come out. And if you play it forever, every possible poem and every possible story will have to come out"

Michael Ende[21]

The development of machine learning models and their embedding in the business bring about a number of benefits that are the result of improved decision-making processes and task automation in model development. These benefits materialize, for example, as a more accurate prediction of demand, improved stock management and pricing strategies, increased customer loyalty, or improved efficiency and lower production costs, among others. This in turn leads to better results in product development or in the provision of services, a more efficient distribution of resources or a better market positioning, which can translate into competitive advantages over competitors who make less use of these techniques.

However, the model building process also has challenges linked to the development and implementation of these new methods:

- On the one hand, machine learning models often require large amounts of data to avoid overfitting, which implies the need to invest in obtaining, ingesting, storing and managing data sources and IT architectures with in-house or cloud solutions to ensure the availability and quality of the data used.

- On the other, it is necessary to invest in model development, model validation, model implementation in the business as usual, as well as in algorithm monitoring and maintenance.

- Likewise, the traceability of the model building process needs to be considered, and the interpretability of the algorithms and the results obtained needs to be ensured, since decisions based on algorithms, even if only partially, should be backed by this knowledge.

- All of the above requires proper governance to ensure both management and ethical issues in the use of models and regulatory requirements are given due consideration. These impacts are even greater for firms operating in regulated industries, since there are limitations to the implementation and use of these models for certain purposes.

- Finally, to comply with the above, it is necessary to have specialist profiles, either by hiring data scientists and developers directly, or by outsourcing the process to specialized companies, as well as to transform the organizational structure and adapt it to accommodate the model development needs of firms, incorporating new ways of working (for example, through Agile organizations[22]).

These challenges have motivated the emergence and development of AutoML systems, since their use can resolve the difficulties encountered, allow firms to take advantage of the benefits of automation and contribute to the democratization of modeling processes, facilitating their use by non-expert users.

Automation rationale

The no free lunch principle[23] suggests there is no model whose performance is always better than all the others, so that, depending on the data set analyzed, the type of model that best predicts or explains the data may be different. Extending

[21] Michael Ende, "The Never Ending Story" (1979). German writer of the 20th century, best known for his fictional and children's works.
[22] The detail of the transformation towards agile organizations has been extensively described in "From Agile delivery, to an agile organization", Management Solutions, 2019.
[23] Wolpert & Macready, 1997.

this idea to the field of machine learning, this principle can be interpreted as the non-existence of estimators or combinations of configurations, hyperparameters or network architectures that are always better than other alternatives. Although academic research shows that selecting values for hyperparameters in small ranges can guarantee optimal models[24] when dealing with specific problems, this idea cannot be extrapolated to all possible problems. This situation means methods need to be found to ensure a machine learning algorithm properly searches for all possible configurations to maximize its performance.

If a problem is to be solved using machine learning techniques, the way to tackle it is to establish the different parameter and configuration options that may be chosen throughout the process. To do this, it is first necessary to identify the characteristics of the data to be processed, as well as the techniques to be used. Then, the modeling approach and the metrics to be used in model selection need to be established and, last, any constraints need to be included based on knowledge of the problem (for example, the sign of the variables). A modeling workflow is then configured that will serve to obtain a group of models ranked according to performance. This flow can then be used to perform an automation operation using a componentization approach, i.e. separating the different model-building processes into components that can be run in a modular way.

Modeling workflow optimization and componentization

There are many options to combine different techniques at each stage of the modeling workflow. Although both the choice of parameters and the configuration of techniques (which ones to apply, in what order, to which part of the dataset, etc.) vary depending on the problem, techniques can be developed to achieve the automation of multiple tasks. Among the main objectives of this automation are to reduce both the cost and the operational errors potentially arising in end-to-end development for each machine learning problem, as well as to improve the efficiency of the modeling process.

Three of the causes behind the need to invest in automation are the following:

- **Redundant development:** some programmable tasks and functions in the modeling process may have been developed in previous processes, either internally by specialist teams in the organization or by the data science community.
- **Errors:** the development of new code usually carries a greater probability of containing errors. This requires a trial and testing process, for which more time and resources are needed.
- **Efficient search for strategies:** finding strategies to explicitly discard configuration combinations and hyperparameter ranges that are considered inadequate or may lead to implementation errors[25].

[24] See, for example, Segal, 2004.

[25] However, although generalizing the problem helps to achieve an efficient resolution, in some cases it is necessary to establish mechanisms so that the modeler can test certain configurations or hyperparameter definitions outside the search space.

Two ways to increase modeling process automation are task componentization and automating the search for optimal configurations and hyperparameters.

Workflow componentization

Splitting the modeling tasks into different parts, and programming and developing them independently, allows modelers to automatically use each part in the form of calls to the developed code, having only to adapt the parameters and configurations, within the possible options, to solve a specific task. This treatment, which is similar to that applied in the development of libraries in programming environments, or to object-oriented programming, allows the task of developing the code, as well as the specific programming language, to be isolated from its subsequent application. This allows for an agile modeling environment. Each of these components receives a certain input (usually a dataset and a set of parameters), and executes a specific task, returning as output another dataset with the result of the applied task.

Modeling process componentization has a number of advantages over developing a workflow for each problem. Some of these advantages are:

Elements of an automated machine learning system

- **Parameter:** an internal property of the model, learned during the learning process and necessary for making predictions.
- **Hyperparameter:** a parameter that cannot be obtained during the process, and must be set in advance. The values that hyperparameters must take to solve a specific problem are unknown. The number of trees in a Random Forest or the number of clusters in a K-Means are examples of hyperparameters.
- **Configuration:** the different value combinations a hyperparameter can take.
- **Configuration/search space:** set of all possible hyperparameter configurations for which an optimal configuration is searched to make the best prediction possible.
- **Neural network architecture:** refers to both the number of layers and neurons present in each layer, and the way in which they are connected. In some cases, the way in which they are trained is also included in the concept.
- **Cost function:** function whose minimum values match the optimal configuration. Searching for the optimal configuration is equivalent to finding minimum values for the cost function. Some cost functions can be the mean squared error or the cross entropy, among other options.

- **Standardization:** in the development of modeling components, which reduces error frequency and improves comparability.
- **Improved quality:** in component development and implementation.
- **Improved efficiency:** in component implementation, in the review by the internal validation and audit areas, and in the approval process.
- **Specialization:** in the development of each component by subject specialists.
- **Improved usability:** in implementation, since these packages can be used by different types of users, including those who do not have a knowledge of programming.
- **Scalability:** in both internal host and cloud-based development, at the subsidiary level and by geographical area.

Automating the search for an optimal configuration

Once the different steps in the process have been separated into components, the best parameters for an optimal process need to be selected for the modeling to be carried out. One approach to selection is to automate the search for parameters using strategies that address this problem systematically, orderly, and leaving a trace of the process created by each possible combination, but at the same time making it possible to assess the impact of decisions made at each stage of the process on final model performance.

However, during the automation process, evaluating the entire set of options is often very complex and is usually conditioned by the limited computing time available due to the number of possible options and combinations, model complexity, and the amount of data that needs to be analyzed. Despite the above, an automated search for configurations, although limited, has some advantages over a manual search, such as:

- **Search optimization:** since it allows for a number of combinations to be produced for later evaluation, and for the best performing combinations to be selected.
- **Problem generalization:** since it makes it possible to create wide search spaces when there is no prior information to anticipate which search subspaces are more likely to generate higher performance models.
- **Search space adaptation:** in problems for which there is some information about what the optimal search space should be, it is easy to adapt this space around constraints in order to improve the results.
- **Experimentation:** since it allows organizations to determine the impact of microdecisions made in each automation component on final performance. For example, it makes it possible to assess changes in different model parameters, such as the maximum depth of trees in a random forest algorithm.

An AutoML system can be defined as a method that allows machine learning models to be built without the need for human intervention and subject to certain computational constraints[26]. The role of model developers in an AutoML system focuses on the choice of data, on selecting the validation criteria for the data, and on choosing the metrics to be used in model assessment and selection, instead of spending their time on processing data and optimizing hyperparameters iteratively based on model results. All this determines the search space, generating a number of options that the algorithm evaluates, subject to the set conditions. This process then results in a number of models ranked according to performance.

The following section will look further into the AutoML approach and its impact when it comes to solving the challenges described in this section. It will also examine the transformation AutoML is bringing to the way models are built and to the entire machine learning workflow.

[26] Yao and others, 2018.

Modeling workflow

How a workflow is defined depends on both the problem to be solved and the type and quality of the data used. There are different methodologies for developing machine learning projects, such as KDD (Knowledge Discovery in Databases), CRISP-DM (Cross-Reference Industry Standard Process for Data Mining) or SEMMA (Sample, Explore, Modify, Model and Assess), among others. Although there are differences between them, there are also some common elements to them all that are key to building machine learning models. The modeling process is summarized below:

Problem identification and planning: this phase involves setting the business objectives and understanding the problem to be solved through a machine learning algorithm, as well as establishing the KPIs that will be used to measure project success. Project planning can then take place.

Data preparation: this phase involves pre-processing of the data, and includes data collection (obtaining, labeling and classification of the data collected and improvement of existing data), cleaning and preparation (processing the data to make it usable), analysis (pattern detection and hypothesis development), visualization (graphical representation to identify trends, outliers or patterns), integration (combining different datasets for a unified vision) and feature engineering (converting raw data to data with the desired shape, creating new variables and selecting variables to include in the model).

Model development: this phase refers to selecting the model and its training, performance evaluation and statistical criteria evaluation, as well as the adjusting of parameters. Model selection involves comparing the different models available to find the one that best fits the data, while training involves comparing the different configurations to convert the information provided into patterns and relationships. Performance evaluation is the analysis of model performance through metrics using data not used during model training, and parameter adjustment involves reviewing the possibility of improving model prediction by readjusting hyperparameters.

Evaluation, validation and approval: this phase involves evaluating whether the business objectives set at the beginning have been achieved, and whether the initial expectations have been met. Also, depending on the governance defined and the model classification (tiering) established in the organization's[27] model risk management framework, there may be additional phases in which the validation and audit teams, independently of model developers, perform a review of the different model aspects (data used, methodology, results, documentation, etc.). This validation can include interpretability[28] techniques in order to understand the underlying relationships that explain the model outcome. Likewise, the approval processes established in the organization's governance frameworks must be carried out. In the case of regulatory models, a final approval process must be performed by the supervisor.

Deployment and embedding in the business as usual: once the previous phases have been completed, the model is embedded in the organization's business processes through its implementation in IT architectures, deployment to production, and periodic monitoring of results.

[27] A detail of these management frameworks can be found in the publication "Model Risk Management: quantitative and qualitative aspects", Management Solutions, 2014.
[28] A detail of these techniques can be found in the iDanae Chair publication: "Interpretability of Artificial Intelligence models", UPM and Management Solutions, 2019.

Machine learning process automation frameworks

"We've Heard that a million monkeys at a million keyboards could produce the complete Works of Shakespeare; now, thanks to the Internet, we know that it is not true"

Robert Wilensky[29]

Once the reasons for the componentization and automation of machine learning workflows and algorithms have been determined, the main question arises as to the best approach to carry it out. Specifically, the following questions need to be answered when it comes to automating the machine learning model development process:

- What are the preliminary steps necessary to prepare the data before the modeling process?
- How should the most appropriate algorithms be selected for the data set to be evaluated?
- How should the search space of possible hyperparameters and configurations be determined?
- Should a particular approach be followed to reduce the size of the search space?

These questions have traditionally been answered by analysts manually selecting the criteria based on a priori judgements and configurations that have worked in the past as well as on trial and error methods, which include a randomness component.

In practice, parameter selection faces a number of challenges, whether development is manual or automated:

- A maximalist approach based on the exhaustive review of all possible combinations is unmanageable in terms of both time and computational resource requirements. Even for relatively small datasets or if search constraints are incorporated based on experience, this task is still unapproachable, which means optimization is not possible in some parts of the process.
- The configurations used depend largely on analysts' a priori judgements and manual adjustments, which makes it necessary to explicitly program a large amount of code. Therefore, the choice and performance of many of the machine learning methods used depend on a large number of decisions about their design made manually or based on previous hypotheses.
- When the task is to produce an evaluation function in order to find the relationship between changes in the hyperparameters and model performance, arriving at this can be very expensive, and sometimes this relationship is not clear or does not allow the analyst to infer from the results obtained.
- This constraint does not have an exclusively global interpretation, since the impact on the loss function from changes in hyperparameters cannot even be adequately inferred locally.
- Optimization cannot be performed directly when datasets are large, as run times are long.

Therefore, although there are incentives to carry out systematic and automatic search procedures, configuring these systems requires solving how to evaluate all possible configurations given the existing restrictions.

In view of this, a current approach underpinning the development of AutoML system components is based on:

1. Automating most of the aspects related to data analysis and pre-processing by creating systems that allow data to be processed and variables to be transformed using the most common manual processing techniques.
2. Creating a search space for possible models and parameters where a set of options can be configured to develop them. Through criteria applied across that space, the best models can be obtained, compared and selected.

[29] Robert Wilensky (1996). Professor at the University of California at Berkeley, School of Information; his main field of research was artificial intelligence.

3. Finally, automating interpretability techniques, although separately from the previous optimization model, so that the reports generated can be more easily understood by the different users.

In short, the aim is to achieve a system that automatically finds patterns in the data, selects a way in which these patterns respond to users' questions, and is capable of adequately explaining the results. This replaces the more complex and less business-related tasks and makes the system accessible for expert roles that are business-oriented rather than data science oriented, ensuring that all processes are carried out efficiently and robustly taking into account both computational and run time constraints. Ideally, an AutoML system should allow the following to be automated:

- Data preparation if there are missing values, outliers, poorly categorized data or data containing errors.
- Combining, reducing, transforming, creating or eliminating variables based on statistical criteria.
- The variable selection process.
- Model selection, trying to avoid both overfitting (a model that fits the training data too closely, distorting the prediction for unknown data) and underfitting (the opposite of overfitting: when a model does not fit the data closely enough as to correctly predict the outcome).
- The explanation to the user about the patterns identified in the data so that they can be understood by a human.

The aim of such a system is to carry out all these processes efficiently and robustly, taking into account computational and run time constraints. There are currently many proposed solutions, including frameworks for centralized, distributed, or cloud-based use. Although the current level of development reached by these approaches means they can now compete with and beat human experts in machine learning, there are still many questions that must be resolved in order for these approaches to be correctly implemented.

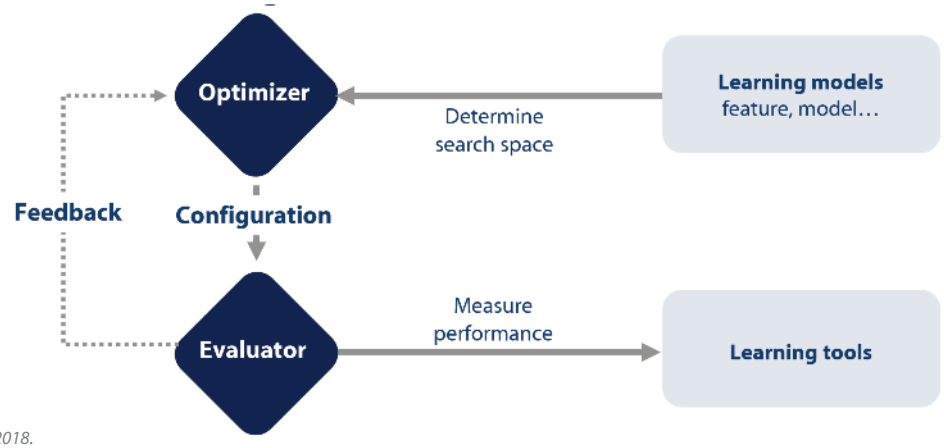
A general framework containing all parts that can potentially be automated is shown in Figure 3[30]. This framework is based on the interaction between two fundamental components: an optimizer, which works in a defined search space, and an evaluator.

On the one hand, the optimizer creates and updates the configurations using a search space determined according to the chosen model and the previously carried out data preparation. The evaluator then measures the performance of the configurations proposed by the optimizer. Depending on the selected approach, the evaluator may affect the optimizer's search strategy.

Generally speaking, the components that are automated within the flow are some of the data preparation stages (preprocessing, feature engineering, processing of missing values, scaling, etc.), the modeling (algorithm selection, hyperparameter optimization, etc.) and, finally, the outcome evaluation. Some data preparation tasks are often left out of the automation process, as they require business insight. In the same way, model interpretability is not automatically assessed, although tools that help to understand the results are usually included.

[30] Yao, y otros, 2018.

Figure 3: general framework for an AutoML system.



Source: Yao and others, 2018.

Although there are many options, it is necessary to develop an AutoML system where the method, described as a theoretical framework, becomes a set of tasks (in the form of programs) that are related to each other (by separating the tasks into components, or through an end-to-end design). Thus, this system consists of a workflow that automates the design of the modeling process.

In general, the market offers mixed approaches where some phases in the AutoML flow (such as data preparation, the processing of variables, and model explainability following model selection) are separated, while the feature engineering, algorithm selection and model evaluation components are included in an optimization model.

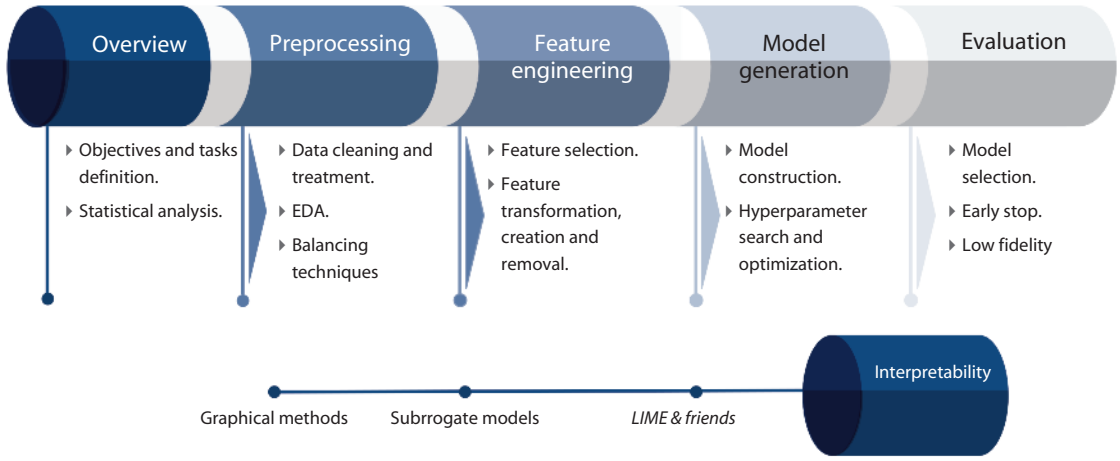
In turn, modeling approaches in AutoML flow design are classified into those that emphasize the modeling process (model-based schemes) and those that focus on data (data-driven approaches). Model-based schemes require a priori knowledge of both the business and the mathematical-statistical component that supports the model, while data-driven approaches are based on processing the data directly, without partitioning by component in the modeling process.

AutoML system components

An AutoML system can be separated into various components, so that, as can be seen in Figure 4, its basic anatomy contains the following modules:

- **Summary:** data set exploration phase that will define the overall set of options for the AutoML process.
- **Preprocessing:** stage involving the cleaning and transformation of raw data before processing and analysis.
- **Feature Engineering:** process that uses the knowledge provided by the data to create variables aimed at improving the performance of machine learning algorithms.
- **Model development:** hyperparameter search and model optimization process.
- **Model evaluation:** implementation of a set of metrics to assess model accuracy.
- **Interpretability:** outcome interpretation using a combination of techniques or models.

Figura 4: autoML system components[31].



Source: He, Zhao, & Chu, 2019.

Hyperparameter optimization

The optimizer uses various techniques to find the best hyperparameter configuration so as to maximize model performance. From a technical standpoint, the role of the optimizer is to search for optimal configurations in the search space to find the global minimum cost function, or at least a local minimum cost function. A distinction can be made between simple techniques (such as grid search, random search, evolutionary algorithms, or Bayesian optimization) and experience-based techniques (such as meta-learning or transfer learning).

For its part, the evaluator uses various techniques to estimate the performance of the configurations proposed by the optimizer, the simplest being to train the model. When this is too computationally expensive, it may be necessary to use subsamples or to include an early stop.

Optimizer techniques: from greedy methods to meta-learning

Once the search space has been set, it is necessary to set an optimizer that searches for configurations in that space. Two of the most common approaches are Grid Search and Random Search, in which no assumptions are made about the search space.

[31] He, Zhao, & Chu, 2019.

A Grid - or brute force - search establishes a grid in the search space, and evaluates the combination obtained at each point on the network. This type of search, which was first applied in an AutoML context in 1990[32], does not necessarily lead to a good configuration (i.e. a local minimum) and can be computationally expensive if there are a large number of hyperparameters. From this approach, others have been developed that improve the process by using an initial grid to explore all the search space regions and later a finer grid in the regions with better performance, with the possibility to iterate the process until a local minimum is found. However, although results are improved, the computational cost of this type of technique remains high.

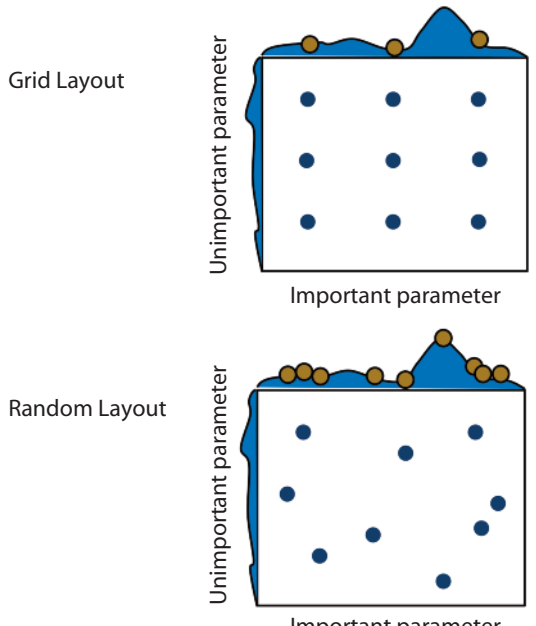
One of the first solutions to improve on the results of a Grid search is Random Search, which is based on selecting a random point in the search space. This allows searches to be made in areas of the search space that are not evenly distributed, thus making it possible to evaluate areas with a higher performance (see figure 5). This technique is still computationally expensive, although as a solution it meets the convergence condition: the longer the search time, the more likely it is to find the optimal set of hyperparameters.

Some of the more elaborate algorithm approaches are the evolutionary algorithms (including genetic algorithms). Fist, these algorithms randomly produce an initial population of configurations. Next, they evaluate the performance of all individuals in the population and select the best performers to produce a new generation based on the former. Furthermore, it is possible to add mutations to the new generations so that they differ from the previous generation. These types of algorithms make it possible to optimize a wide variety of problems, but are also not very efficient in terms of computational cost, since it is still necessary to evaluate all individuals in all generations.

One of the risks of both the grid, random search and evolutionary algorithms is that they can repeatedly explore very low-performance regions in the configuration search space, with no possibility to include a condition in the programming of the algorithm to correct this behaviour. Bayesian optimization (used at least since 2005[34]) solves this problem by creating a probabilistic model of the cost function, through which it selects the best possible hyperparameter configurations in order to evaluate them and estimate the true cost function. Bayesian optimization can update the model iteratively by tracking the results of previous evaluations. This means the probabilistic model can be updated in each calculation.

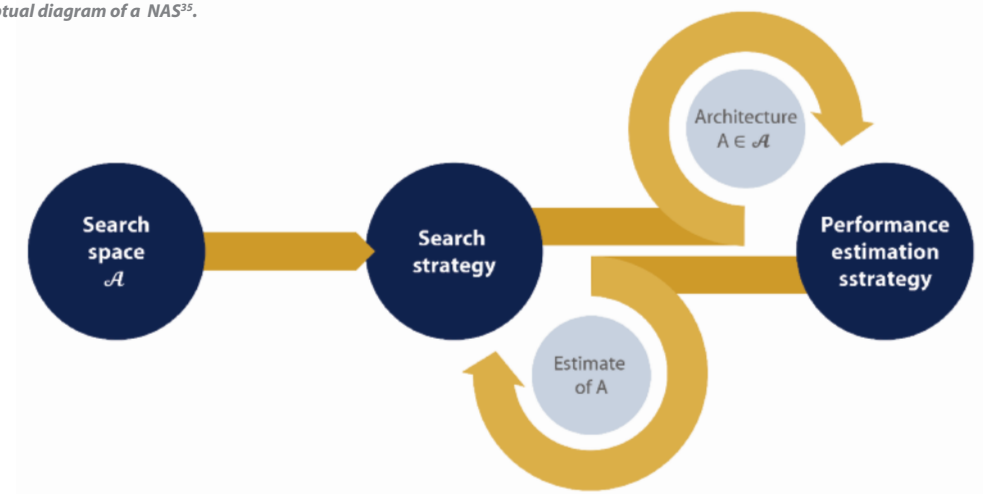
[32] Michie, Spiegelhalter, Taylor, & Campbell, 1994.
[33] Bergstra and Bengio, 2012.
[34] Fröhlich & Zell, 2005.

Figure 5: grid and random search comparison[33] .



Source: Bergstra and Bengio, 2012.

Figure 6: conceptual diagram of a NAS[35].



Source: Elsken, Metzen, & Hutter, 2019.

In some cases the above processes cannot be applied, for example if data are lacking. In other cases, where the datasets may be similar to other previously studied datasets, this knowledge will not be applied. This motivated the development of the meta-learning approach, also known as "learning to learn", which consists of designing machine learning models that are capable of imitating human behavior, quickly learning new concepts and skills using a reduced number of samples, i.e. it aims to design models that can acquire new skills and are able to quickly adapt to new environments based on a few cases.

Evaluator techniques

The simplest way to evaluate the settings provided by the optimizer is to directly evaluate the training and test data. Due to the large number of configurations that the optimizer is expected to provide to the evaluator in an AutoML process, this method can be very time consuming or computationally expensive. There are some approaches to speed up the evaluation process, although this usually means a loss of predictive capacity in the models obtained. These techniques include: evaluating subsets of training data, early stop processes - in which the evaluator stops evaluating if performance is very low in the first iterations, reuse of parameters trained in previous models to initialize the new model and, finally, the use of surrogate models to predict performance, generally using the experience of past evaluations.

Neural Architecture Search (NAS)

As a result of the increasingly wider use of deep learning techniques in areas such as image recognition, voice recognition and machine translation, one of the domains that has attracted the most interested is the configuration of neural network architectures. Similarly to what was previously discussed, these configurations are usually created manually by human experts, which leads to the previously mentioned errors.

As an alternative, neural architecture search (NAS) is based on the use of different techniques to automate neural network design. The aspects on which parameters are set are similar to those previously discussed: search space, search strategy and performance estimation. By using these techniques, the entire process is determined simultaneously, as indicated in Figure 6.

[35] Elsken, Metzen, & Hutter, 2019.

AutoML implementation approaches

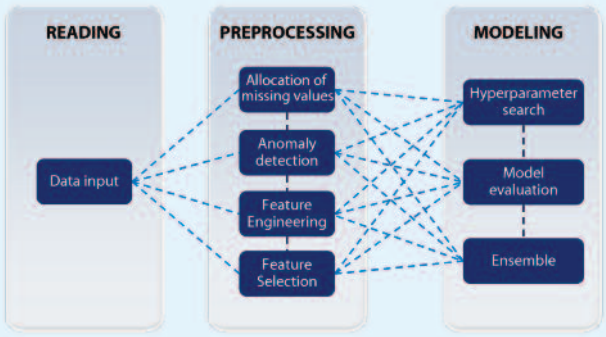
In practice, some considerations need to be taken into account when implementing an AutoML system, such as the system user's profile or the required analysis depth and customization. There are two main implementation approaches:

- One approach is to design a partially or fully editable flow chart (figure 7) for the user to define both the data processing flow and the techniques to be used in each phase of the process. In this case, the level of automation is lower, since it only applies to module execution after the sequence in which models are to be run has been defined. Due to these characteristics, this approach is more suitable when the user has advanced technical knowledge.
- An alternative approach is to use end-to-end automation, with a predefined flow (Figure 8). The data follows a process in which the order of each AutoML component is set according to the general machine learning model building pipeline. This means the user does not have to modify the order in which the development components are run. The user can choose the types of techniques to be applied to each component, but always following the predefined order. These characteristics make this approach more suitable when the user does not have advanced technical knowledge, which is usually the case in business-oriented profiles.

Currently, none of the approaches involves automated development of new variables based on original ones. The reasons are both computational (creating random transformations of variables has a very high computational cost) and business related (expert knowledge of the type of problem being dealt with makes it possible to know which transformation is the most appropriate, and allows experts to better interpret the outcome).

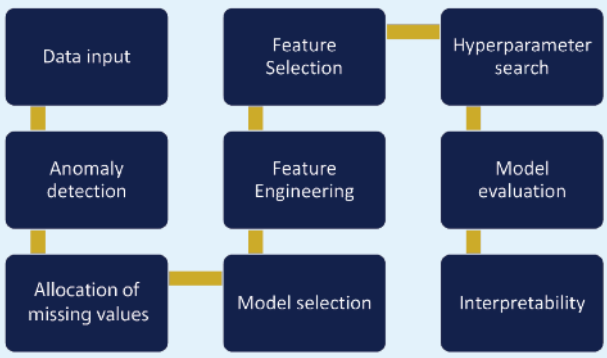
[36] The modeling tool includes a module, Model Creator, which is based on end-to-end automation, and an alternative module, Model Component, which allows the flow to be created by the user.

Figure 7: partially or fully editable flow chart.



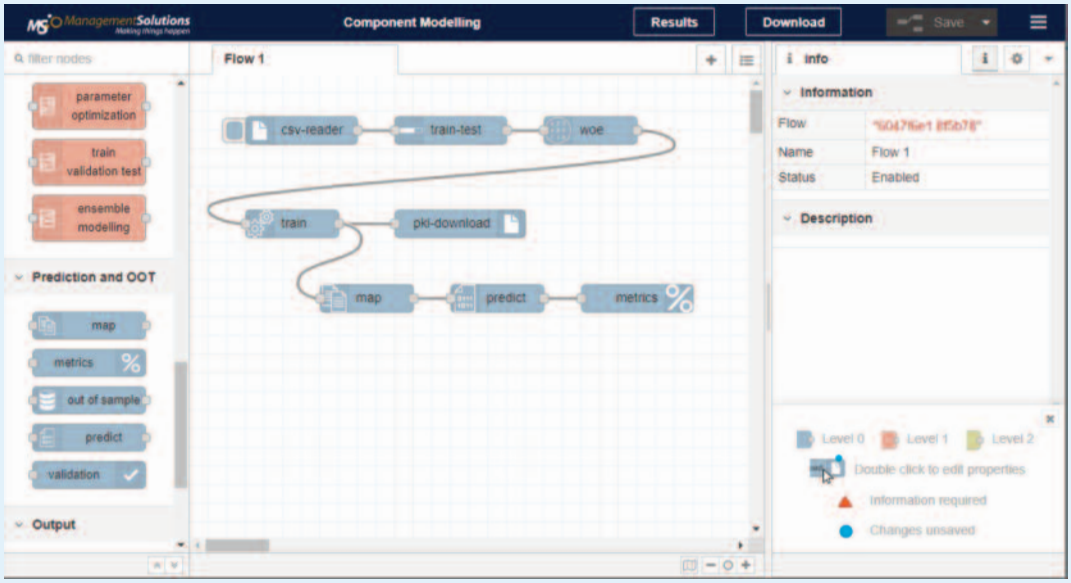
Source: Management Solutions.

Figure 8: predetermined flow.



Source: Management Solutions.

Figure 9: workflow designed in the component-based modeling tool developed by Management Solutions[36].



Source: Management Solutions.

Although when using NAS approaches the elements in the process are determined simultaneously, previous processes, such as those related to data preparation or feature engineering, are usually necessary and their correct implementation results in improved predictive power.

Various elements are included in the search space, such as the number of algorithm layers, the type of operation each layer performs, the hyperparameters associated with these operations - like the number of filters or the size of the associated kernel - as well as the relationship and hierarchy between the different layers depending on the algorithm used. If there is a priori information about possible architectures that usually work properly for a specific task, it is possible to reduce the size of the space, thus simplifying the search. Some methods can also be used to reduce the search space, like setting it through stacks of layers instead of on the entire architecture.

As for search strategies, and as previously mentioned, a number of strategies can be used, including Bayesian optimization, random search and evolutionary methods or methods based on previous experiences such as reinforcement learning (RL). In practice, these techniques do not produce better results compared to the strategies based on random searches. A recent academic study[37] has emphasized reasons such as the use of search spaces being restricted by these algorithms or the impact of the weight distribution across the different layers on the final decision as the elements that limit the results.

Regarding strategies for optimizing NAS performance, the fact that some "low fidelity" approaches use shorter training times, training on subsets of data or fewer filters per layer, may lead to performance being underestimated. Other strategies include extrapolating the learning curve, using surrogate models to predict the performance of new architectures, or initializing the network either by using weights obtained from previously trained networks or through graph theory.

Although NAS has reached a level of performance that can compete with manual configuration, the reasons explaining why the selected architectures work well are unclear at the moment. Similarly, more empirical verification is needed to determine whether the reasons that make a configuration work can be generalized to different problems.

Current challenges in AutoML systems

Currently, and despite the fact that there is still room for improvement, AutoML systems have reached a stage of development where they can compete, and even beat, human experts in machine learning, and have become a key tool that has changed the type of work carried out by the professionals involved. As a result, data scientists are able to focus on the analysis that takes place before and after models are developed, as well as on maintaining methods and systems.

Some of the current challenges are related to improving the process, as well as incorporating interpretability and allowing easier interaction by experts:

- Currently, most innovations are aimed at model selection and optimization, with less attention on data processing and preparation. This is due to the difficulties encountered in automating some of the processes without incurring a high computational cost.

[37] Sciuto, 2019.

- Another open question is how to deal with elements that have very low interpretability, known as black boxes, since incorporating them in decisions can lead to legal, ethical and technical problems. Some of the main lines of research in this regard are focused on Explainable AI, interpretability, and improving model traceability and transparency. Although these concerns are also shared in the traditional way of developing machine learning models, the greater automation offered by AutoML systems means there is more emphasis on their use in these processes.

- On the other hand, for AutoML systems to be effective, they must allow users to interact with them and modify and overwrite their decisions. This is so users can incorporate expert business knowledge on different aspects of the process, such as the predictions made or the complexity and interpretability of the models.

- Last, there is a need to establish benchmarks for comparing performance between the different solutions proposed, as well as for the metrics used to measure this performance to be clearly defined.

Augmented machine learning

An approach that is attracting a great deal of attention due the generalized use of AutoML methods and systems is the so-called Augmented Machine Learning. Under this approach, some processes are automated so that AutoML systems can deal with the complexity arising from the increase in possible architectures, hyperparameter options and training options, but there is still an expert that uses the outcome from the tool and evaluates the alternatives and combinations produced taking a holistic view. There are several reasons why this approach is used:

- The first is that these systems cannot incorporate the context information that the user has on the data, therefore the process seems to improve when the user guides the system in the search for data patterns. This concept, known as representation engineering, is, for instance, common in areas such as the interpretation of internet searches[38].

- Furthermore, siloed information analysis through these tools reduces the value expected from the use of advanced analytics, which means that a data science expert has a key role in making decisions on questions such as when the different data sources should be combined or in which cases transfer learning techniques should be used, since AutoML systems cannot at present analyze all possible options before making a decision.

- Finally, and as mentioned in the previous point, the ethical questions related to the goal, the data used, and the potential bias generated in the decision process, make it necessary for analysts to assess both whether a model should be used in a decision-making process, and the limitations of the model. Generally speaking, using models as support in decision making works well, while automated judgment, though improving, is imperfect. When it comes to predicting human behavior, the outcome it produces is questionable[39].

[38] Abbasi, Kitchens, & Ahmad, 2019.
[39] Narayanan, 2019.

AutoML competitions: a tool for exploring AutoML approaches

"Ford!, he said, there's an infinite number of monkeys outside who want to talk to us about this script for Hamlet they've worked out"

Douglas Adams[40]

As we have seen in previous sections, a specific approach is still not favored over others in areas such as data processing prior to modeling, how to select algorithms, or how to configure them properly. However, some trends and standards are beginning to emerge that should be incorporated into any process involving Advanced Analytics techniques, in particular an AutoML system.

A common way of evaluating different AutoML approaches has been to organize competitions between data scientists with the aim of building AutoML systems. These provide a good benchmark given that they allow the different approaches to be compared under equal conditions and, therefore, make it possible to extract from their outcome whether there are some preferable configurations and under what circumstances they work best. Initially, these competitions were aimed at evaluating the choice of models and hyperparameters[41]. Later, this type of exercise was refined, with participants now expected to develop an automated and computationally efficient system capable of training and evaluating models without any human intervention[42].

Broadly speaking, the main aim of these competitions is to i) find out the effect of time constraints on algorithm design, ii) identifying which tasks are more difficult and for which type of participants , iii) find out if there are specific configurations that tend to work better for specific types of datasets or problems, and iv) evaluate the impact that optimizing hyperparameters and configurations has on final model performance.

A review of AutoML competitions

A number of patters emerge from the analysis of different competitions[43]:

- In general, the use of heuristic approaches or of grid or uniform searches on a linear or logarithmic search space is common.
- In some cases, the above method is improved through the use of regularization.
- Overtraining is controlled by including stopping criteria in iterative optimization methods.
- The separation between the training and validation data sets is not usually optimized.

As a result, it can be seen that in no case is the entire process automated, and human intervention is needed in exercise definition-related tasks. It is still difficult to select a system to suit a specific type of problem and to adapt it to the existing data set.

Regarding the variables, the difficulties encountered are directly related to some attributes found in the analyzed data sets, such as the existence of unbalanced data, scarce data, missing values or categorical variables. In these cases, greater intervention is required to identify, treat and assess the impact of such treatment on the process.

As for the configuration and hyperparameter selection process, the main problems stem from ad hoc techniques that worsen model performance, such as the use of unsophisticated techniques to separate the training and testing samples, inappropriately selecting model complexity, hyperparameter selection considering only the test sample, not using all computational resources, or inappropriately defined performance metrics.

From the point of view of search methods, many are based on either grids or uniform distributions over the parameter search space, although there are some sophistications that build on regularization methods, or Bayesian approaches that incorporate stopping conditions to avoid overfitting.

Management Solutions AutoML competition

[40] Douglas Adams, "The Hitchhiker's Guide to the Galaxy" (1979). English writer and screenwriter, best known for the saga of the same name.
[41] Véase, por ejemplo, NIPS 2005.
[42] NIPS 2016, ICML 2016 y PAKDD 2018.
[43] Hutter, Kotthoff, & Vanschoren, 2019.

Aim and definition

In a similar spirit, Management Solutions designed an internal competition aimed at producing an AutoML algorithm capable of making predictions in different datasets without modifying the code, with a time limit to encourage computational efficiency The proposed exercise was based on the use of supervised approaches to solve binary response problems under the following conditions:

- 3 datasets with different sizes (<100 kb, <1 Mb and <5 Mb), all of them consisting of a balanced sample.
- No missings values, with both categorical and continuous variables, and including irrelevant variables.
- Computational resources limited to: Windows 10 computer, Intel Core i5-6300 CPU@ 2.40GHz 2.50GHz processor and 8 Gb of RAM, and a maximum runtime of 20 minutes for each dataset.

Evaluation

The submitted function was evaluated on three different datasets, similar to those submitted as training samples. For this, the following aspects were taken into account:

- Area under the curve (AUC) metric (50%).
- Quality and cleanliness of the code and use of the PEP8 standard (20%).
- Use of Object Oriented Programming (10%).
- Originality (20%).

Results

More than a hundred professionals from a number of backgrounds like physics, mathematics, engineering and economics, entered the competition as part of more than seventy teams in total. Many of the participants have - or are studying for - a postgraduate degree in data science. They are from very diverse geographical areas: Peru, Chile, Colombia, Brazil, Germany, the United States and Spain.

Throughout the competition, participants faced various choices regarding data processing, model development and parameter optimization. Most teams prepared the data by removing possible outliers and correlated variables, normalizing variables, reducing dimensionality and allocating missing values. Some teams used WOE or one-hot-encoding techniques for categorical variables and specific processing for any unbalanced data sets, and considered interactions between variables to increase predictive capacity, or the removal of irrelevant variables such as constant variables, with very little variance, or categorical variables with a very large number of categories with respect to the total number of entries.

The aim of this pre-processing is clear: on the one hand, it prepares the data so it can be correctly read by the models

used, and, on the other hand, it reduces search space dimensionality so that less time is required to find an optimal configuration. Other participants took a different approach to dealing with this problem, limiting the number of algorithm runs to a specific and constant number, or limiting the number of models that are evaluated by the system.

In most cases, hyperparameter optimization was approached using grid searches. Some teams used random search, genetic algorithms, or Bayesian search. A participant notably implemented a random search routine to subsequently perform a search in a small area around the optimal configuration found. This was aimed at improving the metric with the resulting configurations in the event that the outcome from the random search did not exceed a specific score.

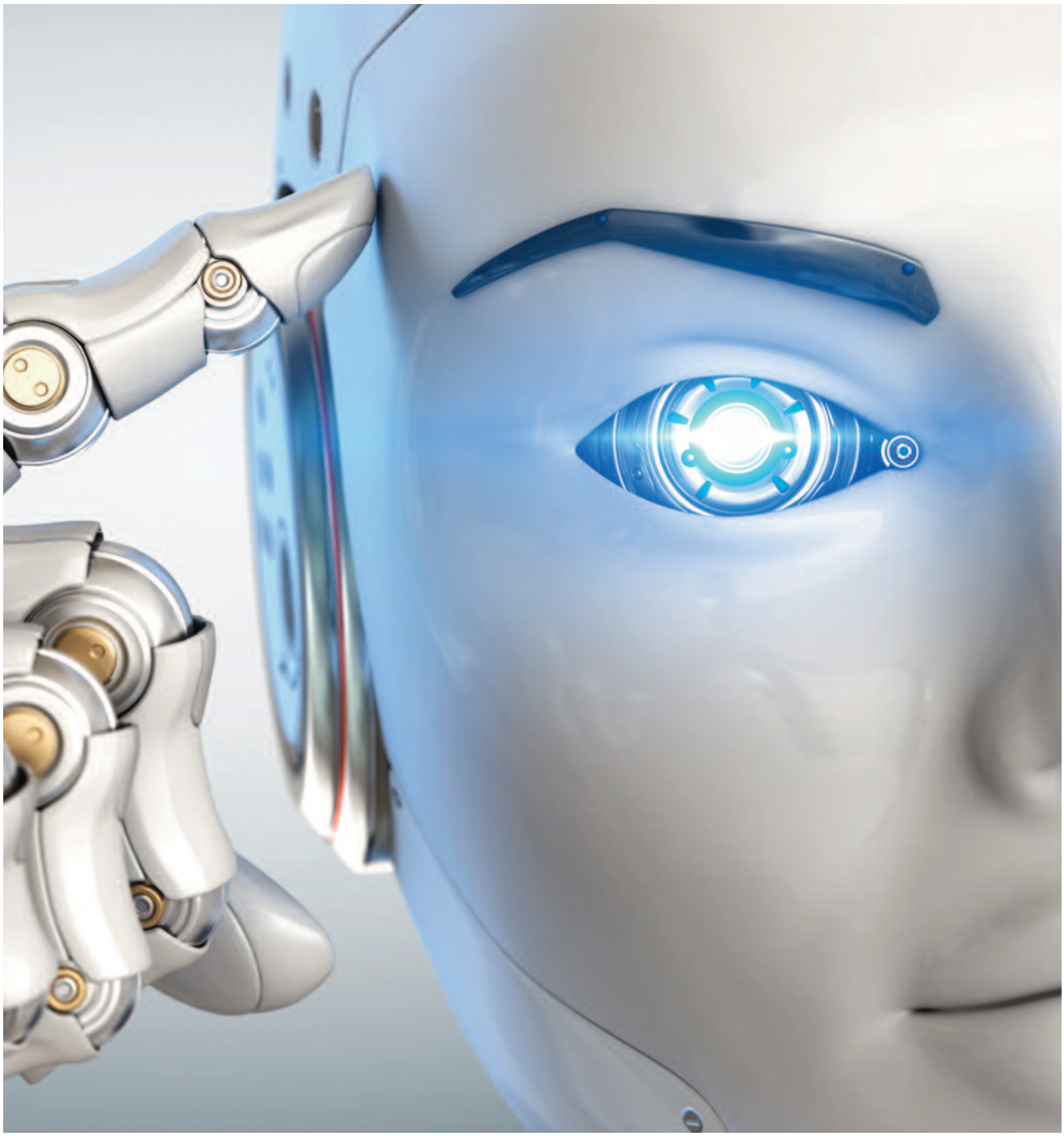
The teams used cross validation to assess the performance of the proposed configuration, and implemented models mostly obtained from the scikit-learn library, with some exceptions where keras, lightgbm or xgboost were used. To optimize computational time, some participants carried out a previous study of the most predictive variables to work only with them, while others evaluated a list of models and stopped evaluating when the maximum defined time was reached, hence there may be models estimated but not evaluated in the sample.

In general, the work carried out included optimizing the entire pipeline, using stacking models or task parallelization at various points. Some participating teams also included interpretability modules, either to interpret datasets or to interpret aspects of the AutoML process such as the choice of using a model over another.

The time limit to run each dataset did not have a great impact in general, since the evaluation files were small and the partipants' AutoML systems did not have any problems running the files within the set time. Only some participants limited the number of models to be evaluated in order to avoid exceeding the stipulated time.

Final insights

Well? What do you think of my new poem?
I once read that, given infinite time, a thousand monkeys with typewriters would eventually write the entire Works of Shakespeare
But what about my poem?
Three monkeys, ten minutes
– Scott Adams[44]



Current situation and challenges of AutoML

The configurations used to obtain machine learning models depend significantly on the a priori judgement and manual adjustments of analysts. This means that, to develop models using machine learning techniques, it is necessary to explicitly program large amounts of code. The choice and therefore the performance of many machine learning methods depends on a large number of decisions about their design, made manually or based on previous hypotheses. This may result in suboptimal values being chosen, which may in turn lead to overfitting in models developed with small datasets and underfitting in larger datasets[45], indicating that improvements are still required to ensure these systems are properly used in industry.

While AutoML approaches have reached a level of development that can compete with and beat human experts in machine learning, there are still many issues that need to be resolved in order for them to be applied correctly. The main challenge facing current AutoML systems is for design decisions to be made with a data-driven approach, objectively and automatically.

In any case, the above is not incompatible with the user having the possibility of interacting with the system and being able to modify and overwrite the decisions it makes. In this sense, machine learning model development is a craft industry where experts tackle problems by designing manual solutions. These solutions are in many cases developed "ad hoc" for a specific project, and incorporate the data scientist's preferences and a priori judgements, often leaving out the sensitivity of business experts. An understandable AutoML system interface for business analysts avoids manual decisions in configurations, and at the same time makes it possible to incorporate business decisions on issues like the sign or importance of the variables, the choice of models based on the interpretation of projections, the sensitivity to scenarios, or the complexity and interpretability of the models obtained.

Another open question is how to deal with black box elements that limit interpretability and can lead to legal, ethical and technical problems if they are incorporated into decision making. Some of the research in this connection is notably aimed at Explainable AI, interpretability, and improving model traceability and transparency.

Finally, some aspects like search process efficiency are constantly being improved, as different competitions have shown[46].

Level of development

Advances in AutoML are happening at different paces: most innovations are aimed at some feature engineering and model selection techniques, while there is still a long way to go in the area of data preprocessing and preparation[47]. This has an impact on the types of tasks that need to be carried out in organizations, as well as on the level of related employment.

On the one hand, AutoML replaces the more complex and less business-related tasks in the design of pipelines for specific problems, allowing business-oriented roles with less technical knowledge of machine learning to design the entire pipeline.

On the other hand, it requires an infrastructure that enables these processes to be carried out and be properly updated, either through outsourced AutoML services or through an in-house AutoML system with specialized teams that ensure the workflow runs correctly.

Furthermore, some issues have hardly been addressed by AutoML systems, so tasks such as data integration or cleaning, or the development and maintenance of variables, in addition to some machine learning aspects such as unsupervised learning or reinforcement learning, are not usually integrated into these systems.

AutoML systems are expected to become a key tool that can modify the type of work that needs to be carried out, thus freeing up data scientists to spend their time on analysis tasks required both before and after models are built, on developing AutoML systems, and on solving problems where generic AutoML tools do not allow for adequate configuration.

[44] From a Dilbert comic strip by Scott Adams.
[45] For example, in the case of HPO methods. See Hutter, Kotthoff, & Vanschoren, 2019.
[46] Hutter, Kotthoff, & Vanschoren, 2019.
[47] Ibid.

Bibliography

Abbasi, A., Kitchens, B., & Ahmad, F. (2019). The Risks of AutoML and How to Avoid Them. Harvard Business Review.

Bank of England. (2019). Machine learning in UK financial services. Bank of England.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of machine learning research.

Cátedra iDanae. (3T-2019). Interpretabilidad de los modelos de Machine Learning. Cátedra iDanae.

Cátedra iDanae. (4T-2019). Ética e Inteligencia Artificial. Cátedra iDanae.

CrowdFlower. (2017). Data Scientist Report. CrowdFlower.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural Architecture Search: A Survey. Journal of Machine Learning Research.

European Banking Authority. (2020). EBA report on Big Data and Advanced Analytics. European Banking Authority.

European Comission. (2020). White paper on Artificial Intelligence - A European approach to excellence and trust. European Comission.

Fröhlich, H., & Zell, A. (2005). Efficient Parameter Selection for Support Vector Machines in Classification and Regression via Model-Based Global Optimization. IEEE Xplore.

Gartner. (2019). How Augmented Machine Learning Is Democratizing Data Science. Gartner.

He, X., Zhao, K., & Chu, X. (2019). AutoML: A Survey of the State-of-the-Art. arXiv preprint arXiv:1908.00709.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated Machine Learning: Methods, Systems, Challenges. Springer.

Management Solutions. (2014). Model Risk Management: Aspectos cuantitativos y cualitativos de la gestión del riesgo de modelo. Management Solutions.

Management Solutions. (2018). Machine Learning, una pieza clave en la transformación de los modelos de negocio. Management Solutions. Obtenido de Management Solutions.

Management Solutions. (2019). De proyectos Agile, a organizaciones Agile. Management Solutions.

Michie, D., Spiegelhalter, D., Taylor, C., & Campbell, J. (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Narayanan, A. (2019). How to recognize AI snake oil.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development. IBM J. Res.

Sciuto, C. &. (2019). Evaluating the Search Phase of Neural Architecture Search. Sciuto, Christian & Yu, Kaicheng & Jaggi, Martin & Musat, Claudiu & Salzmann, Mathieu.

Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression.

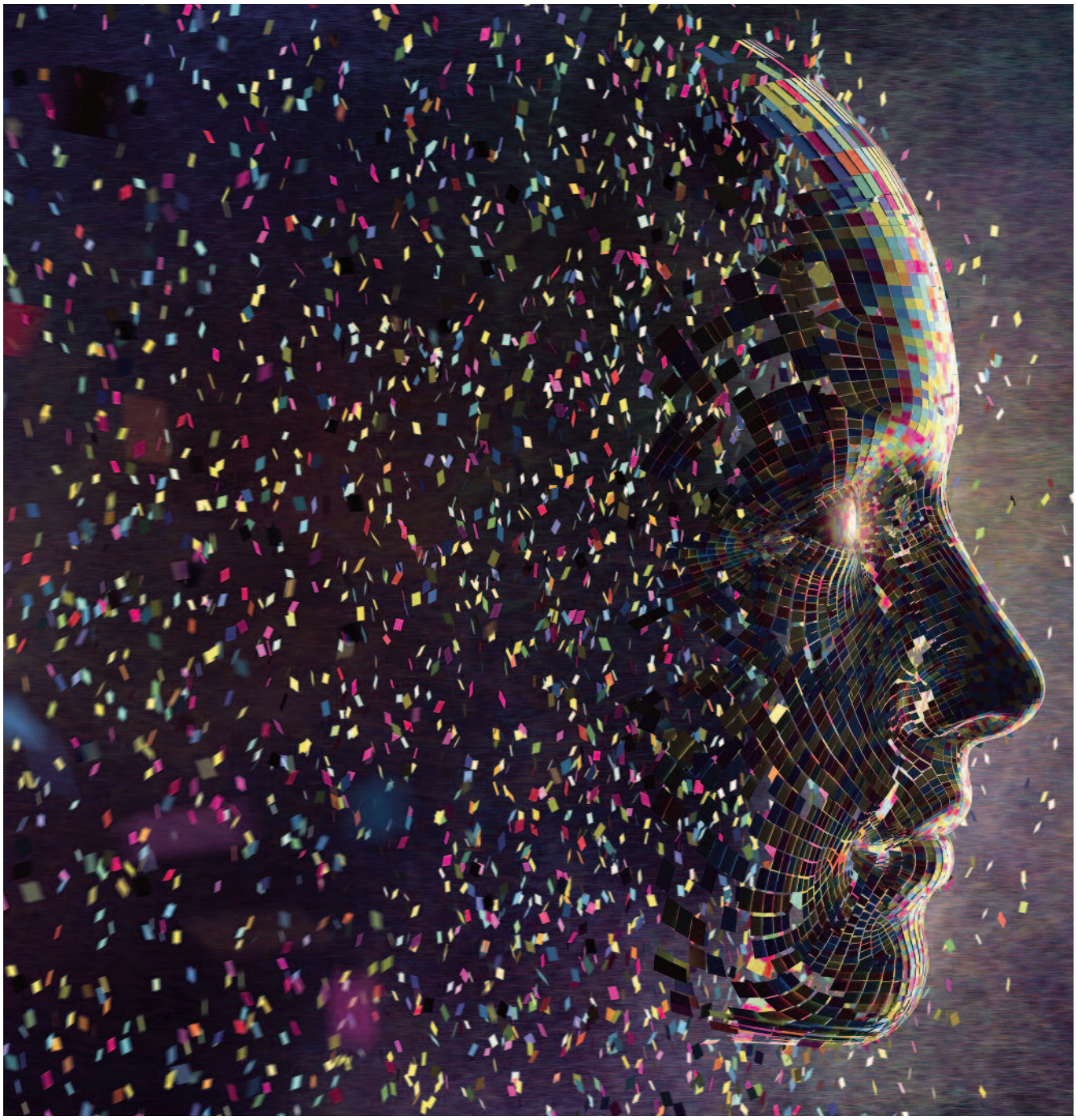
Stanford University. (2019). AI Index Report. Stanford University.

Statista (2019). Cost decreases from adopting artificial intelligence (AI) in organizations worldwide as of 2019, by function. Statista.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. IEEE transactions on evolutionary computation.

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., . . . Yu, Y. (2018). Taking Human out of Learning Applications: A Survey on Automated Machine Learning. arXiv preprint arXiv:1810.13306.

Glossary



Cloud Computing: type of computing that makes resources such as data storage or computational power available without active management by the user.

Configuration: the possible combinations of values that hyperparameters can take.

Continuous / Categorical Variable: a continuous variable is a numeric variable that can take any value between two limit values. A categorical variable can be a discrete numeric variable, or it can be words or another type of variable.

Correlated variables: variables that have a similar behavior.

Cross Validation: a technique that consists of dividing a sample into k groups, and iteratively using each of group for validation and the rest for construction, changing the validation group in each iteration.

Configuration space: the set of all possible configurations on which the optimal configuration is searched to make the best prediction possible.

Cost function: a function whose minimum values match the optimal configuration. Searching for the optimal configuration is equivalent to finding minimum values for the cost function. Some cost functions can be the mean squared error or the cross entropy, among other options.

Dimensionality reduction: a process by which the search space is made smaller, either through a combination of variables, elimination or other methods.

Early Stop: a technique that consists of stopping the search process earlier than planned if certain requirements are met.

Feature Engineering: process of extracting features from data using data mining techniques and specific domain knowledge.

Grid / random / evolutionary / bayes / meta / transfer: different methods used to search for hyperparameter optimizations.

Hyperparameter: a parameter that cannot be obtained during the process and must be previously set. The values that hyperparameters must take to solve a specific problem are unknown.

Machine learning: a field of computer science that focuses on developing techniques that enable a program to learn to find patterns in a data set.

Metric: measure to assess the performance of a model.

Missing values: missing values within a dataset.

Normalization: data preprocessing to make the mean value of a variable close to zero and between -1 and 1.

Outliers: values that, due to having been poorly measured or being atypical, are numerically far away from the rest of the data.

Overfitting / underfitting: characteristic of a model that occurs when the model fits the training sample too closely / not closely enough, so that satisfactory results are not obtained on samples other than the training sample (e.g. on the validation sample).

Parameter: internal property of a model that is learned during the learning process and is necessary for making predictions.

Regularization: a mathematical technique that consists of adding a component to the cost function to detect those variables that are not contributing significantly different information to the model. It is used to avoid overfitting problems (such as in the case of elastic nets).

Surrogate model: a generally simpler model that tries to emulate a more complex model in certain environments or situations.

WOE (Weight of Evidence): data processing for categorical variables

Our aim is to exceed our clients' expectations, and become their trusted partners

Management Solutions is an international consulting services company focused on consulting for business, risks, organization and processes, in both their functional components and in the implementation of their related technologies.

With its multi-disciplinary team (functional, mathematicians, technicians, etc.) of 2,500 professionals, Management Solutions operates through its 31 offices (15 in Europe, 15 in the Americas and 1 in Asia).

To cover its clients' needs, Management Solutions has structured its practices by sectors (Financial Institutions, Energy, Telecommunications and other industries) and by lines of activity (FCRC, RBC, NT), covering a broad range of skills - Strategy, Sales and Marketing Management, Risk Management and Control, Management and Financial Information, Transformation: Organization and Processes, and New Technologies.

The R&D department provides advisory services to Management Solutions' professionals and their clients in quantitative aspects that are necessary to undertake projects with rigor and excellence through the implementation of best practices and the continuous monitoring of the latest trends in data science, machine learning, modeling and big data.

Javier Calvo Martín
Partner at Management Solutions
javier.calvo.martin@msgermany.com.de

Manuel Ángel Guzmán
Director R&D at Management Solutions
manuel.guzman@managementsolutions.com

Daniel Ramos García
Supervisor R&D at Management Solutions
daniel.ramos.garcia@managementsolutions.com

Segismundo Jiménez
Supervisor R&D at Management Solutions
segismundo.jimenez@managementsolutions.com

Carlos Alonso Viñas
Consultant at Management Solutions
carlos.alonso.vinas@msspain.com

Management Solutions, Professional Consulting Services

Management Solutions is an international consulting firm whose core mission is to deliver business, risk, financial, organization, technology and process-related advisory services.

For further information please visit **www.managementsolutions.com**

Follow us at: 

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña London Frankfurt Paris Amsterdam Copenhaguen vOslo Warszawa Zürich Milano Roma Lisboa Beijing New York
Boston Atlanta Birmingham Houston San Juan de Puerto Rico San José Ciudad de México Medellín Bogotá Quito São Paulo Lima Santiago de Chile Buenos Aires