AI Decision Coordination: easing the appropriation of decision automation for business users

AI Decision Coordination : faciliter l'appropriation de la décision automatisée par les utilisateurs métier

Thomas Baudel, Grégoire Colombet, Raphael Hartmann
IBM France Lab
Orsay, France
baudelth@fr.ibm.com

ABSTRACT

In a previous article [4], we proposed a methodology, ObjectivAIze, to objectivize the implementation of automated business decision systems, that replace or complement analysts in routine decision tasks: processing alerts, validating application files... This methodology relies on the definition of metrics of performance: costs associated to a correct decision, an incorrect decision, a human intervention and allocation of tasks to humans or machines based on their relative performance. This methodology has been implemented for clients in the banking industry, through consulting missions. To facilitate the appropriation of this methodology by business users, who know the business but are less familiar with statistics and process modeling, we are developing an analysis and decision support software, AI Decision Coordination. This article presents the very first version of the software, its design choices, the reaction of the first users as well as the future planned evolutions.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Pointing**; **Visualization techniques**; **Empirical studies in HCI**; • **Applied computing** → *Business process management*.

IHM'23, April 03–07, 2023, Troyes, France

KEYWORDS

Decision Support Systems, Human-AI collaboration, Human-centric AI.

RÉSUMÉ

Dans un article précédent [4], nous avons proposé une méthodologie, ObjectivAIze, permettant d'objectiver la mise en oeuvre de système de décision métier automatisée: traitement d'alertes, validation de dossiers d'application... Cette méthodologie repose sur la définition de métriques de performance: coûts associés à une décision correcte, à une décision incorrecte, à l'intervention d'un humain; utiliser ces métriques pour allouer la prise de décision à la machine ou l'humain en fonction de leur performance relative. Cette méthodologie a été mise en oeuvre chez des clients dans l'industrie bancaire, par des missions de conseil, les rendus étant essentiellement sous la forme de rapports de recommandations présentés aux responsables métier. Afin de faciliter l'appropriation de cette méthodologie par ces responsables, qui connaissent le métier mais moins les statistiques et la modélisation de processus, nous élaborons un logiciel d'analyse et d'aide à la décision, AI Decision Coordination. Cet article présente les choix de conception, la toute première version du logiciel, l'accueil des premiers utilisateurs ainsi que les évolutions à prévoir pour les satisfaire.

MOTS CLÉS

Processus métier, aide à la décision, collaboration humain-algorithme.

INTRODUCTION

Business processes increasingly incvolve algorithmic decision aids: when a decision has to be made - a purchase, a quotation, a recruitment or a diagnosis... - the software can use the results of previous similar decisions to suggest a choice, including a confidence score. If the confidence score is high, the process can probably be fully automated. If not, it can help human analysts to ensure the consistency of decisions. A priori, these systems should improve decision-making processes, enhancing their reliability and performance, either by freeing humans from repetitive, low-value-added tasks, or by enabling processing far greater volumes of information. However, the possibility of effective human-algorithm collaboration is far from being systematically confirmed by the literature [1].

A major obstacle remains: when a decision task is fully or partially automated (decision support), who bears responsibility for the individual decisions made? The analysts who developed the model, while able to certify certain properties of the algorithms and learning data, cannot assume full responsibility, as they do not control its deployment and usage. Project owners, who oversee the decision-making process, are placed in a delicate position: while they understand the requirements of the process and can accept an algorithmic implementation, they face difficulties in identifying

its potential limits and the possible interactions between human decision-making and algorithmic recommendations. In short, it is difficult for them to take ownership of a solution whose operation is largely beyond their control. Furthermore, if the algorithmic decision replaces human decisions taken by collaborators, it is no longer possible to invoke a certain sharing of responsibilities between supervisors and supervisees: a heavy weight, moral if not regulatory, is concentrated on a small number of people. The result is that, according to market surveys [6] and our own experience, a large proportion (85%) of decision automation projects are not implemented.

The aim of the present work is to propose solutions to implement algorithmic decision-making with confidence, and thus enabling business managers to appropriate algorithmic decision tools through tangible, controllable and supervisable evidence, expressed in the vocabulary of the business rather than that of the data analysis engineer.

METHODOLOGY

In a previous article [4], we proposed a methodology based on an empirical approach to address this issue: ObjectivAIze. Its principles, briefly recalled here, consists in:

- Define decision metrics: benefits of a good decision, costs of a bad decision, cost of human expertise... This cost model makes it possible to assign a performance to each decision. It is developed collectively, iteratively involving all the stakeholders of the process: data scientists, process designers, process owners, management, compliance officers...
- Compare, through experimentation, the decisions made by algorithms and analysts in several configurations, on a sufficiently large and representative dataset. This common practice is called a "parallel run".
- Partition the space of decisions made along the axes for which we observe a significant difference in performance between algorithms and human agents.
- On each element of the partitioning obtained, allocate decision making to the agent (human or algorithmic) that maximizes performance.

A preferred partitioning axis is the algorithm's confidence score, taken as the best approximation the algorithm can provide of the probability of correct classification. There is strong experimental evidence for this:

- when the probability of success of a classification task is high (typically above 80%), the use of algorithmic recommendation by an analyst, who confirms or invalidates the proposed choice, tends to degrade overall performance [9].
- below a threshold of around 70%, on the contrary, algorithmic recommendation is counter-productive [8]: low algorithmic confidence reflects a lack of information in the available data, which the analyst will have to compensate for by acquiring other information.

Figure 1: Comparison of algorithmic, human and augmented decision processes

- Finally, in [5], we showed, on a specific task, that it was possible, on segment 70%-80%, to make collaboration productive.

Our methodology, ObjectivAIze, can be described as an application of Fitts' function allocation theory [10], aiming to determine the functions for which the machine overperforms the human or vice versa. It is summarized in figure 1.

On this idealized diagram, we identify the confidence score ranges on which to allocate the decision to the algorithm (0-20 and 80-100), to the human alone (40-60), or to the human with an algorithmic recommendation (20-40 and 60-80). We applied our methodology to several decision tasks: international financial sanctions alert filtering, transfer fraud alert filtering and anti-money laundering, and each time found comparable results, with vaiours nuances that we won't detail here.

BUSINESS NEEDS

Our methodology has been implemented via consulting missions: a team of consultants collects, analyzes and interprets the results, creating various cost and workload scenarios, then delivers its conclusions to the project owner, in a report that we hope will be didactic, but which ultimately remains the product of an analysis by specialists acting on behalf of the process owner.

Our presentation, with supporting data and illustrations, helps convincing project owners, who appreciate the inclusion, through the cost model, of their business considerations for the deployment of algorithmic decision solutions. This model also provides the opportunity to discuss alternative hypotheses, and to plan process reorganizations that go further than simply introducing an algorithm to support human decisions.

However, we are convinced that the key to adopting our methodology in more varied settings lies in the process owner's ownership of the decision allocation strategy. To assume responsibility for delegating decisions to algorithms, the process owners must be empowered to performn the allocation themselves, with measures he or she understands, and legible feedback on the impact of choices made, in a process dashboard. It is with this in mind that we are developing IBM AI Decision Coordination, marketed in December 2022. We propose a tour of its main functionalities, it being understood that the entire value chain presented has not yet been fully developed.

AI DECISION COORDINATION

Dans le logiciel, chaque projet décrit une tâche de décision. Chez nos clients actuels, c'est une même équipe qui supervise un ensemble de tâches de décisions dévolues à des chargés de clientèle, prenant à leur compte tout ce qui concerne la gestion d'un portefeuille de clients: détection des alertes sur transactions inhabituelles (de plusieurs types), réponse à des requêtes... Cette équipe est en charge de tout le soutien aux chargés de clientèle dans leurs nombreuses tâches. Elle supervise donc un certain

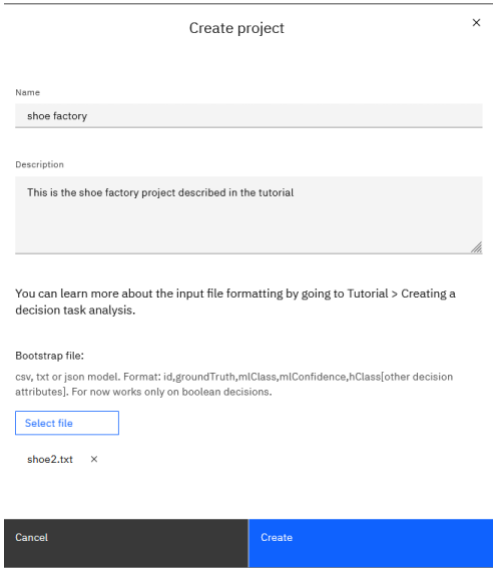

Figure 2: Initializing a project from parallel run data

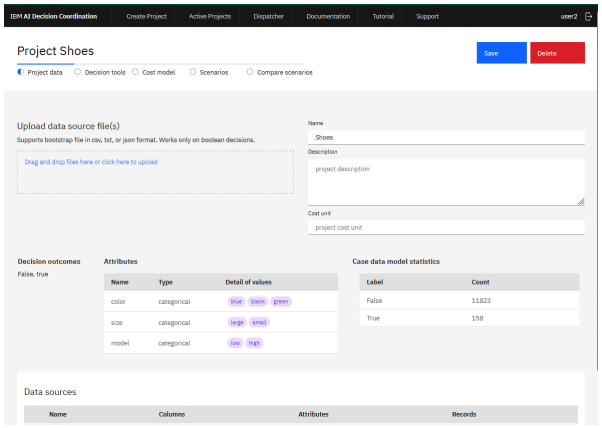

Figure 3: Main presentation page

nombre de tâches de décision, chacune ayant son cycle de vie propre. Pour illustrer la présentation, nous supposons que la tâche étudiée est un processus de filtrage d'alertes fraude: un premier algorithme filtre, parmi les millions de transactions quotidiennes, celles ayant des caractéristiques suspectes. Si une alerte est levée, elle est envoyée au chargé de clientèle, qui pourra contacter le client pour autoriser ou interdire la transaction. L'alerte peut également être propagée à une cellule spécialisée. La tâche du maître d'ouvrage est de déterminer à partir de quels seuils de risque une alerte vaut la peine d'être remontée, sachant qu'un grand nombre de faux positifs va dégrader la faculté d'attention des humains et agacer les clients par des blocages intempestifs.

In the user interface, each project describes a decision task. With our current customers, the same team supervises a set of decision tasks assigned to employees, who are manage a set of common tasks for a customer portfolio: detecting alerts on unusual transactions (of various types), responding to queries... This team is in charge of supporting the account managers in their many tasks. The team oversees a number of decision-making tasks, each with its own lifecycle. To illustrate the presentation, we'll assume that the task under study is a fraud alert filtering process: a first algorithm filters out, from the millions of daily transactions, those with suspicious characteristics. If an alert is raised, it is sent to the account manager, who may contact the customer to authorize or close the transaction. The alert can also be forwarded to a specialized unit. The task of the project manager is to determine the risk thresholds at which an alert is worth escalating, bearing in mind that a large number of false positives will impair the vigilance of the account managers and irritate customers with untimely interruptions.

Creating a project

The project lifecycle begins at its creation: description of a decision task, important attributes of the decision (origin and recipient, amount, various features), the possible outcomes (reject, put on hold, authorization...), and an initial data set: past results with the performance of humans, of the algorithm we wish to deploy, and a truth value, which, in the case of fraud, has the advantage of being knowable a posteriori.

This information is generally available and prepared by the team of data analysts who developed the machine learning model: indeed, a classifier cannot be developed without this data. To simplify initialization, all the user has to do is load an initial analysis file in the format provided (essentially column names): the structure of the decision task will be inferred immediately, which is a major gain in efficiency (figure 2).

The project is immediately usable with standard performance and statistical metrics giving an idea of raw performance and estimated human-algorithm complementarity (figure 3). We allow the user to create several automated decision implementation scenarios, which will enable comparing several deployment solutions under consideration, with different cost and capacity assumptions.

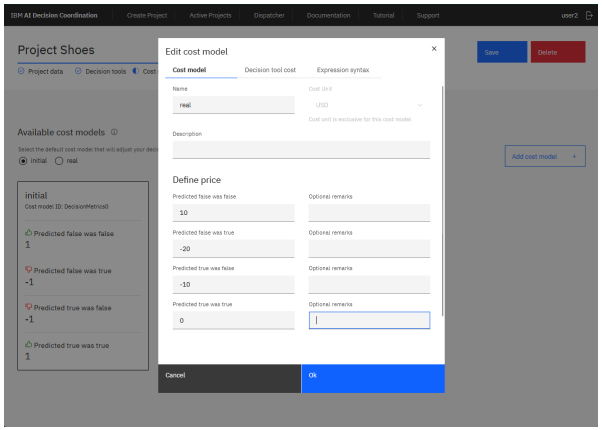

Figure 4: Definition of performance metrics (cost model).

Defining a performance model

L'étape suivante consiste à définir un ou plusieurs modèles de coût réalistes, c'est à dire reflétant la réalité de l'organisation humaine dans laquelle s'insèrent les décisions (Figure 4). Pour cela, à chaque combinaison de "décision prise/bonne décision", on associe une formule de coût. Par exemple, pour un virement frauduleux, une décision de blocage rapportera 0, tandis qu'une autorisation à tort coûtera le montant de la transaction. On défini également des coûts et capacités associés à la décision: coût du travail humain, chargé de clientèle ou expert, chiffrage des éventuels désagréments subis par le client. Les formules utilisent un langage proche des macros d'excel, familières à nos utilisateurs principaux. Bien évidemment, la mise au point du modèle de coût pour un processus de décision est itérative et co-construite par différentes parties.

Un des grands avantages de notre proposition est de permettre la discussion autour de différentes hypothèses, en visualisant rapidement leur impact potentiel avant tout déploiement. Par rapport aux rapports que nous rendions précédemment, c'est un progrès important pour permettre l'appropriation de la configuration proposée par ceux qui en sont ultimement responsables.

The next step is to define one or more realistic performance models, that reflect the reality of the human organization in which the decisions are made (Figure 4). To achieve this, we associate a cost formula to each dimension of the confusion matrix. For example, in the case of a fraudulent transfer, a blocking decision will yield 0 (no loss, no gain), while authorizing a fraudulent transaction will cost the amount of the transaction. We also define the costs and capacities associated with the decision: the cost of human resources, whether customer service representatives or experts, and the cost of any inconvenience to the customer. The formulas use a language similar to Excel macros, familiar to our main users. The development of the cost model for a decision-making process is iterative and co-constructed by different parties.

One of the great advantages of our proposal is to enable discussing different hypotheses and quickly visualizing their potential impact before any deployment. Compared with our previous reports, this is a major step forward in ensuring that the proposed configuration is appropriated by those who are ultimately responsible for it.

Function allocation

Finally, for each scenario, an optimal allocation strategy is proposed, in several forms (figure 5):

- Key indicators: average cost per decision, margins of error, distribution of decisions between humans and automation...
- Summary graph showing proposed distribution of decisions.
- Textual description, in the form of a decision tree, of the proposed allocation strategy.

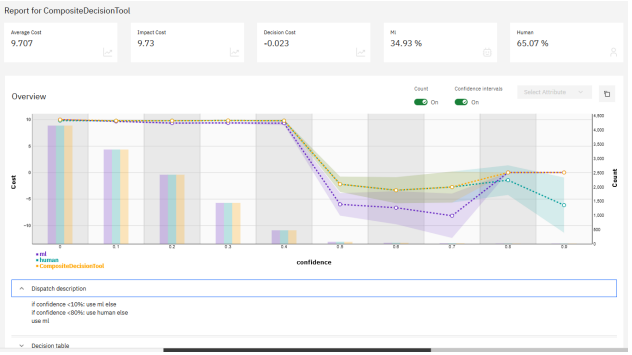

Figure 5: Proposed function allocation

- Detailed graphs showing, for each level of the decision tree, allocation strategy and key indicators.

The proposal is editable: it is possible to change the proposed thresholds, to evaluate the impact of threshold changes on the proposed cost model: this reverse calculation should enable cost assumptions to be revised, when the inferred solution proves unrealistic in practice. For example, estimating human costs that are too low would lead to a saturation of processing capacities, showing that either we have been unrealistic, or that a serious recruitment effort would be justified to improve overall performance. Techniques similar to those we have implemented for editing data from visualizations [3] and [2] will be useful for this.

Once an allocation strategy has been agreed upon, it is marked as ready for deployment: it will be archived, and implemented in the customer's transaction processing systems by the information systems service. As of now, we don't consider a process manager will deploy an allocation strategy directly: it has to be tested and validated by a specialized team before being put into production.

Eventually, however, it will be necessary to provide monitoring capabilities, to reinforce the client's confidence in his allocation choices. This will require further development and extensive integration into the user's information systems, which will require further work.

FIRST REACTIONS

To date, our customers have been convinced by the clarity of our methodology and its ability to integrate business constraints and priorities into the deployment of an automated decision-making solution. When we deploy our solutions, however, it is necessary to share them with data analysts, who demand to retain control over subjects they consider to be their attribution. To this end, the ability to copy/paste all our analyses into flexible tools such as Jupyter Notebook is much appreciated. It means we don't have to draw up an ever-growing inventory of useful features to provide, at the risk of losing novice or occasional users.

It's easier to convince decision-makers and business managers of the merits of the solution we offer, even if it requires more extensive training than we'd like, to enable a manager to carry out his or her own analyses. We're leaning more towards making the interface more didactic, to guide them towards simple presentations, at least at the outset, even if a great deal of customization work needs to be carried out by an experienced analyst with a good understanding of the subject.

CONCLUSION

Increasing regulatory and legal constraints are weighing on the use of automated decision-making. For example, Article 14 of the draft European regulation on AI, [7], specifically addresses the issue of human control ("Oversight") and requests the implementation of specific measures to preserve human

autonomy. Our methodology and the software that supports it should enable organizations to benefit from the phenomenal potential of algorithmic decision-making, by objectivizing the overall process performance (human, technological and organizational) rather than focusing on the ML models as most solutions currently provide.

ACKNOWLEDGEMENTS

The authors would like to thank the design and development team: Ollie Day, Rongron Zhu, Varun Sharma, Ali Siddiqi, Kevin Quinn, Ali Gondal, Philippe Charman, Frédéric Delhoume, and our management: Steve Astorino, Angel Montedescoa, Namik Hrle and Amélie Hocquette for believing in the project.

REFERENCES

[1] Eugenio Alberdi, Lorenzo Strigini, Andrey A. Povyakalo, and Peter Ayton. 2009. Why Are People's Decisions Sometimes Worse with Computer Support?. In *Computer Safety, Reliability, and Security*, Bettina Buth, Gerd Rabe, and Till Seyfarth (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 18–31.

[2] Thomas Baudel. 2002. *A Canonical Representation of Data-Linear Visualization Algorithms*. Technical Report. ILOG Research Report. arXiv:cs.GR/1412.4246 https://arxiv.org/abs/1412.4246

[3] Thomas Baudel. 2006. From Information Visualization to Direct Manipulation: Extending a Generic Visualization Framework for the Interactive Editing of Large Datasets. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST '06)*. Association for Computing Machinery, New York, NY, USA, 67–76. https://doi.org/10.1145/1166253.1166265

[4] Thomas Baudel, Grégoire Colombet, and Raphaël Hartmann. 2022. ObjectivAIze : Déterminer empiriquement les rôles respectifs de l'humain et de l'algorithme dans la prise de décision. In *33e conférence internationale francophone sur l'Interaction Humain-Machine (IHM'22)*. AFIHM, Namur, Belgium. https://hal.archives-ouvertes.fr/hal-03835662

[5] Thomas Baudel, Manon Verbockhaven, Victoire Cousergue, Guillaume Roy, and Rida Laarach. 2021. ObjectivAIze: Measuring Performance and Biases in Augmented Business Decision Systems. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 300–320.

[6] Cem Dilmegani. 2023. *4 Reasons for Artificial Intelligence (AI) Project Failure in 2023*. Technical Report. AI Multiple. https://research.aimultiple.com/ai-fail/

[7] Directorate-General for Communications Networks Content and Technology. 2021. *Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. European Commission, Bruxelles. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206

[8] Linda Onnasch. 2015. Crossing the boundaries of automation—Function allocation and reliability. *International Journal of Human-Computer Studies* 76 (2015), 12–21. https://doi.org/10.1016/j.ijhcs.2014.12.004

[9] Sandra Dorothee Starke and Chris Baber. 2020. The effect of known decision support reliability on outcome quality and visual information foraging in joint decision making. *Applied Ergonomics* 86 (2020), 103102. https://doi.org/10.1016/j.apergo.2020.103102

[10] J. C. Winter and D. Dodou. 2014. Why the Fitts List Has Persisted throughout the History of Function Allocation. *Cogn. Technol. Work* 16, 1 (feb 2014), 1–11. https://doi.org/10.1007/s10111-011-0188-1