IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 February 2025, accepted 25 February 2025, date of publication 4 March 2025, date of current version 14 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3547439

RESEARCH ARTICLE

Explainable Predictive Factors for Inter-Agency Partnership Success

NICOLA DRAGO[1,2], GONÇALO AFONSO BRAZ[3], AND TULLIO VARDANEGA[1], (Member, IEEE)
[1]Brain Mind and Computer Science (BMCS), Human Inspired Technology Research Centre, University of Padova, 35131 Padua, Italy
[2]United Nations University Operating Unit on Policy-Driven Electronic Governance, 4810-225 Guimarães, Portugal
[3]Department of Informatics, University of Minho, 4710-057 Braga, Portugal

Corresponding author: Nicola Drago (nicola.drago@phd.unipd.it)

This work is part of a PhD project, which received a grant from the Department of Mathematics of the University of Padova, Italy.

ABSTRACT Development partnerships that overcome the intrinsic limitations of individual organizations are essential to achieve the goals of the 2030 Agenda for Sustainable Development. However, one in five strategic partnerships fails, incurring higher costs for lesser impact. Assuming that partnership performance is proportionate to co-financed project ratings, we used the corresponding evaluation reports and project ratings from multiple multilateral organizations to automate the identification and analysis of critical predictive factors for partnership success. This approach ultimately helps form more selective partnerships. Our automation engine integrates text mining, sentiment and semantic analysis, and supervised Machine Learning with explainable and opaque algorithms. To achieve this goal, we compared four alternative Machine Learning models. By applying Shapley's additive explanation values to trained models, we ranked factors by criticality, identifying their Top-10 rank. With bootstrap sampling, instead, we computed confidence intervals. Our rating predictions were rather good. Explainable AI models achieved a Matthews Correlation Coefficient of 0.749. Opaque algorithms reached 0.780. Both compare well with the current success rates of 81% for project co-financing partnerships and 75% for knowledge partnerships. Our prototype engine may help organizations improve partnership appraisal efficiency and improve collaboration. Future work will include refining factor formulations, deploying models with incremental learning, and introducing causality analysis.

INDEX TERMS Explainable algorithms, failure factors, machine learning, strategic partnerships, success factors, sustainable development, organizations, semantics, investment, predictive models.

I. INTRODUCTION

A. CONTEXT AND OBJECTIVES

Developmental partnerships are relevant because they synergize political, technical, and financial resources to implement the 2030 development agenda and its Sustainable Development Goals (SDGs). The quest to address the SDGs involves yearly investments in the region of 4.3 trillion United States Dollars (USD) [1], [2], far beyond the reach of any individual organization's political, technical, and financial resources. Strategic development partnerships are, therefore, a necessary instrument to that end.

However, 19%[1] to 25%[2] of existing developmental partnerships are deemed *unsuccessful* as they do not fully achieve their objectives [3], reducing the efficiency and impact of public resource expenditures on beneficiary communities. Estimates based on data from the Asian Development Bank (AsDB) Independent Evaluation Department (IED) in [3] suggest that the budget assigned to partnerships that did not achieve their objectives was USD 10.9 billion in 15 years

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

[1]co-financing partnerships.
[2]inter-agency coordination partnerships.

for AsDB alone. Identifying the most promising partners or partnerships is a pressing concern.

Currently, research on partnership performance focuses almost exclusively on success factors, relying on qualitative, somewhat subjective assessments and overlooking detrimental factors [4]. The underlying problem is the lack of evidence-based knowledge of the factors that underpin or undermine partnership performance.

The work presented in this study aims to fill this knowledge gap. We propose a quantitative method, based on identified factors of influence, to make the selection of strategic partners more objective, targeted, and actionable and to improve the success rate of partnerships from its current range of 75%-81%, depending on the type of partnership.

Two actions enable factor selection based on criticality and usage in partnership performance predictions. The first action is factor identification and visualization to allow factor comprehension and synoptic detection of factor patterns, overlaps, and redundancy. This action is based on a systematic review of the (predominantly gray) literature on partnership evaluation, text mining, semantic analysis, and dimensionality reduction. Its definitions, methods, and results are discussed in a separate publication [5]. The second action is the automatic inference of the factor criticality and related confidence intervals from project evaluation documents and project ratings. These documents and ratings are processed by explainable and opaque algorithms for text preparation, embeddings, sentiment analysis, Machine Learning, and feature selection. The gray literature pertains to the architecture of the method and then to the description of the theoretical framework and the context of partnerships. The peer-reviewed literature is especially relevant for selecting the algorithms used in this method and, to a lesser extent, its architecture. This study presents the outcomes of the second action. The resulting output provides organizations with an evidence-based prioritized list of factors capable of predicting partnership performance, offered for use in the design of future partnerships.

B. DEFINITIONS

This study uses domain-specific terms, briefly defined below.

Partnerships are collaborative relationships between multiple entities working toward shared objectives through a mutually agreed division of labor. Partnerships involve shared responsibilities for outcomes, distinct accountabilities, and reciprocal obligations [6]. In this study, entities are multilateral development organizations because they have a larger geographic scope and project evaluation report repositories than other organizations. There are three types of partnerships: co-financing, coordination and cooperation, and knowledge and learning [3], [7], [8]. This study focuses on project co-financing partnerships between two or more multilateral agencies with a developmental mandate.[3]

This work analyzed projects financed principally by AsDB, the International Fund for Agricultural Development (IFAD), the United Nations Development Programme (UNDP), and the World Bank (WB). Such partnerships typically include national governments and may or may not include other partner types such as bilateral cooperation agencies, international non-governmental organizations, civil society organizations, academia, and the private sector. Non-multilateral organizations fall outside the scope of this study. This study does not consider trust funds or supply agreements as partnerships because they do not entail a division of labor or shared responsibility for the outcomes. Public-Private Partnerships (PPP), a type of project co-financing partnership aimed at resourcing large, costly infrastructure or massive services, are not the focus of this study, which analyzes the broader spectrum of developmental project co-financing partnerships. AsDB, IFAD, UNDP, and WB were informed of this study, formally and informally, by the United Nations University Operating Unit on Policy-Driven Electronic Governance (UNU EGOV) and the University of Padova (UniPD).

According to AsDB [3], *partnership success* is the attainment of "*most or all*" of the objectives of the partnership. Conversely, a partnership is unsuccessful if the objectives are "*achieved only to some extent or not at all.*" According to Liebenthal et al. [6], *"In the end, partnerships should be judged by their results."* Hence, this study measured the success of a partnership based on its overall project rating. Successful partnerships are those with a "highly successful" rating, which corresponds to grade four on a four-grade scale, and grade six and the upper scale of grade five on a six-grade scale. In contrast, unsuccessful partnerships are rated as "highly unsuccessful" or grade one on the same scale. In this study, project rating is the outcome to predict.

A *factor* is a feature of an inter-agency partnership that affects the performance of the partnership. Some examples of the 753 textual factors, manually drawn by the first author from the literature, are in the *factor text* column of Table 8. In addition to textual factors, this work gathered other numeric or categorical project factors: co-financing status,

[3]African Development Bank (AfDB), Asian Infrastructure Investment Bank, Caribbean Development Bank, Council of Europe Development Bank, Development Bank of Latin America and the Caribbean, European Bank for Reconstruction and Development, European Investment Bank, Food and Agricultural Organization of the United Nations, Inter-American Development Bank (IADB), International Fund for Agricultural Development (IFAD), International Labour Organization, Islamic Development Bank, New Development Bank, United Nations Capital Development Fund, United Nations Development Program (UNDP), United Nations Educational, Scientific and Cultural Organization, United Nations Entity for Gender Equality and the Empowerment of Women, United Nations Environmental Program, United Nations Human Settlements Programme, United Nations Industrial Development Organization, United Nations International Children's Emergency Fund, United Nations Population Fund, United Nations Relief and Works Agency for Palestine Refugees. International Bank for Reconstruction and Development, International Development Association, and International Finance Corporation are part of the World Bank Group.

implementation country, project end year, main financier organization, and project code. In this paper, *Top-10* denotes the topmost critical factors in our classification.

In this study, a *simulation* refers to the evaluation of Machine Learning (ML) performance in predicting project ratings using unique combinations of the following components: algorithms (including those for embeddings, sentiment analysis, and ML models), sentence filtration parameters (such as sentence length thresholds, polarity thresholds, inclusion or exclusion of stopwords, and semantic similarity ratio), ML hyperparameters, and the dimensionality of the factor subset (comparing the complete set to a critical subset). Each simulation computes standard ML metrics, focusing on the Matthews Correlation Coefficient (MCC) because of its effectiveness in assessing multiclass classification performance with imbalanced datasets. These simulations aimed to compare project rating prediction performance (as measured by MCC) between the critical factor subset and the complete factor set and between explainable algorithms (logistic regressions, random forests) and opaque algorithms (extreme gradient boosting, multilayered perceptron neural networks). This comparison aims to determine whether explainable algorithms and the critical factor subset can achieve project rating predictions comparable to or exceeding the current partnership success rate, thereby enhancing the transparency and interpretability of predictive models in inter-agency partnership evaluations.

Regularization refers to techniques applied during model training to prevent overfitting. These improve the ability of the ML model to generalize well to new, unseen data rather than fitting too closely to the training data [9].

Appendix contains a glossary of terms and abbreviations used in this work.

C. ASSUMPTIONS

This work makes four assumptions:

1) Project co-financing, a partnership type, is assumed to represent all inter-agency partnership types despite their differing mandates. This is a convenient assumption because projects are endowed with evaluation documents and ratings, which the other partnership types are not, thus denying ML-based assessments.
2) Partnerships between multilateral organizations are assumed to represent all inter-agency partnerships.
3) The co-financed project results are assumed to represent partnership results. The project results are external to the partnership but embed the internal workings of the partnership, such as the relations between the partners, relevance, effectiveness, and efficiency [6].
4) Partnership success is assumed to be proportional to the overall project rating. As a compound index based on expert assessment of multiple project and partnership indicators [10], [11], the project rating condenses the evaluation of the performance of the project and the agency.[4] Project ratings are assumed to provide the most objective and available measure of partnership success and results, which would otherwise be nuanced and sometimes vaguely defined.

D. RESEARCH QUESTIONS

We transform the problem described in Section I-A, centered on the knowledge gap on the factors underpinning and undermining partnership performance, into the following three specific Research Questions (RQ):

RQ1 How can we rank factors after their importance for the success or failure of partnerships?
RQ2 What is the uncertainty in the values that rank the factors by criticality?
RQ3 Can a subset of critical factors be selected to make partner selection more targeted and actionable?

The ultimate goal of the encompassing research project is to answer **RQ3**. Achieving this would allow deploying decision-making software that predicts the success rate of a future partnership based on ex-ante documents, such as feasibility studies, draft memorandums of understanding, and geographical and chronological contextualization data.

E. PAPER ORGANIZATION

The remainder of this paper is organized as follows. Section II reviews the state of the art in partnership performance prediction, project rating prediction, and algorithm usage applied to development cooperation projects. Section III describes the approach used in this study and the inputs used in its implementation. Section IV examines the inputs used to implement the chosen approach. Section V discusses the scripts and models used to select the factors by importance and evaluate the effects and uncertainty of this factor selection. Section VI presents the results of the factor ranking by criticality and related confidence intervals, including the effects of the parameters of the method. Section VII condenses the limitations and potential application of the method in practical settings. Section VIII summarizes the key findings and novelty of this study and proposes potential applications to further ambits.

II. STATE OF THE ART

In 2014, the United Nations published the report ''A World that Counts: Mobilising the Data Revolution for Sustainable Development'', advising the preparation of machine-readable data for a forthcoming data revolution [12]. The AsDB published a comprehensive thematic evaluation report on their partnerships and analyzed six influencing factors [3]. The WB developed a project outcome prediction tool targeting projects rated moderately satisfactory or higher,

[4]Relevance, effectiveness, efficiency, sustainability, and development impacts, and the performance of the agency, the co-financiers, the reference government agency, and the executing agency according to the AsDB [10]. Project effectiveness, efficiency, relevance, outcome, sustainability, monitoring and evaluation system, implementation and execution according to UNDP Independent Evaluation Office (IEO) [11].

which flags at-risk projects using project indicator-based methods, achieving 68% accuracy in ex-post outcome rating predictions [13]. Neither method used ML. Since then, project evaluation report analytics automation has become a trending research topic. In 2018, the United States Agency for International Development, a bilateral development agency, published guidelines on ML and Artificial Intelligence (AI) use in development cooperation, advising on social risks, data quality and preparation, model selection and evaluation, context relevance, and inclusion [14]. In the same year, WB published a comparison of 10 ML classification algorithms–including Logistic Regressions (LR), Random Forests (RF), eXtreme Gradient Boosting (XGB), and multi-layered perceptron Neural Networks (NN)–, Latent Dirichlet Allocation (LDA) topic modeling applied to 145,000 project documents, and a proposal for expanding the corpus to other multilateral development banks [15]. Still in 2018, a keynote talk at the WB on ML and causality and causal forests [16], considered ''a rapidly advancing area with relatively few practical applications to point to'' at that time [17]. In 2020, IFAD published two studies on learning from project evaluation documents and the organization's data to inform future project design, utilizing a range of semantic analysis, and supervised and unsupervised ML methods such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Stochastic Gradient Boosting Machines, Bidirectional Encoder Representations from Transformers (BERT), Word2Vec [18], [19]. McNamara et al. [4] qualitatively focused on the failure factors of partnerships, highlighting the lack of systematic research on the causes of partnership performance. In 2022, AsDB classified social media feedback on their interventions using Valence Aware Dictionary and sEntiment Reasoner (VADER) and SentiWordNet for sentiment analysis and Naive Bayes (NB), SVM, Decision Trees, and RF for classification [20]. That year, UNDP published its Artificial Intelligence for Development Analytics (AIDA) system, enabling the search and retrieval of evaluative evidence from UNDP's 6,000-project-evaluation database, including paragraph-level polarity scores [21]. UNDP AIDA utilizes Amazon Web Service Textract packages to process documents into topically organized paragraphs [22]. Most recent works date to 2023, when the WB created a taxonomy of factors influencing project success, utilizing sentiment and semantic analyses on evaluation report paragraphs with NB, RF, SVM, NN, LDA, and Word2Vec [23]. The WB also published its results, including the distribution, frequency, and sentiment of factors affecting implementation by multiclass outcome rating; it used LR, KNN, SVM, decision tree, RF, NB, and stochastic gradient for classification, and TextBlob, VADER, FLAIR, DistilBERT, and SieBERT for sentiment analysis [24]. In September of the same year, Kumar [25] published a doctoral thesis that used the XGBoost classifier to predict potential partnership opportunities for enterprises and SHapley Additive exPlanation (SHAP) values to enhance the interpretability of the model by highlighting the impact of each input feature on the prediction, which are similar tools to this work but different domains and purposes. Searching the literature with ''partnership,'' ''performance,'' and ''prediction'' returns mainly papers on PPP, with little focus on project co-financing in general.

To date, no semi-automated quantitative study of developmental partnerships and their influencing factors is available in the literature. This study fills the gap in strategic partnerships by providing semi-automated multiclass rating prediction for co-financed projects (partnerships) and non-co-financed ones, SHAP-based factor prioritization for partnership success or failure, and opaque and explainable models that weigh both positive (w_i_+) and negative (w_i_-) contents, starting from Portable Document Format (PDF) documents.

III. METHOD

The method described in this paper identifies critical factors by evaluating and benchmarking multiple simulations. Critical factors are ranked at the top of the simulation that achieves the highest overall project rating (proRate) prediction MCC, evaluated for both the opaque and eXplainable Artificial Intelligence (XAI) algorithm chains. The top factors by SHAP value for co-financed projects rated highly successful (or coFin $= 1$ and proRate $= 4$ following the abbreviations in Appendix) are the partnership critical success factors. Symmetrically, those of highly unsuccessful co-financed projects (proRate $= 1$) are the critical failure factors. Factor ranking quality was measured using confidence intervals calculated using bootstrap sampling.

Thirteen topical simulations were performed to detect the effects of sentence filtration parameters, including the retention of negative and adversative stopwords in sentence preparation, algorithm type (whether explainable or opaque), the factor sub-set dimension, and the utilization of geographical and chronological contextualization values.

To explain how this method ranks the factors, this study analyzed four key simulations and outlined the salient outcomes of the other four simulations as summarized in Table 1.

An 11-step Python script processes a corpus of multilateral organizations' project evaluation reports in PDF format into factor ranking by the SHAP value (Table 2). The core methodological point is the counting of corpus sentences that exhibit both semantic similarity with individual factors and a positive or negative polarity score (see the script's Step 6 for sentence BERT -sBERT- embeddings and Step 7 for Term Frequency - Inverse Document Frequency (TF-IDF) embeddings in Table 2). Individual factors are then characterized by two integer weights, derived from the aforementioned counts: w_i_+ and w_i_-, where w stands for weight, i denotes the $i$eth factor, + denotes positive polarity scores, and - indicates negative polarity scores. The w_i_+ and w_i_- aim to capture the evidence-based positive

TABLE 1. Simulations by type of algorithms, factor sub-set, sentence polarity threshold (neuTres), sentence similarity ratio (simRatio), minimum sentence length (senLeng), and inclusion of negative and adversative stopwords in text. The simulations with the factor datasets limited to 24 and 10 selected factors are intended to evaluate the method's performance using the complete factor set and smaller, less computationally costly subsets of critical factors, for which data gathering is easier. The summary of results by simulation is reported in Table 11.

Topic	Simulation ID	Data file code	Algorithms	Factor sub-set	Parameters			
					neuTres	simRatio	senLeng	stopword
Parameter effects	Simulation 01	sim-21	opaque	753	0.4	0.3	10	Yes
	Simulation 02	sim-38	XAI	753	0.6	0.3	10	Yes
	Simulation 03	sim-01	opaque	753	0.6	0.5	20	Yes
	Simulation 04	sim-42	XAI	753	0.1	0.1	10	Yes
	Simulation 05	sim-41	XAI	753	0.4	0.3	10	No
Algorithm effects	Simulation 01	sim-21	opaque	753	0.4	0.3	10	Yes
	Simulation 04	sim-42	XAI	753	0.1	0.1	10	Yes
Factor selection	simulation-06	sim-37	opaque	24	0.4	0.3	10	Yes
	Simulation 07	sim-43	opaque	10	0.4	0.3	10	Yes
	Simulation 08	sim-45	XAI	10	0.1	0.1	10	Yes

(+) and negative (-) aspects of a project, as reflected in critically evaluated lexicon-balanced reports. These factors render the partnership viewpoint, as they are taken from the partnership literature and focus on partnership performance. The *w_i_+* and *w_i_-* values are normalized by column using *MinMaxScaler* across all records.

Each record, which denotes an individual project, is therefore characterized by a series of 753 *w_i_+* and 753 *w_i_-*, a proRate outcome value, a coFin co-financing status, and a couCode-proYear geographical and chronological contextualization feature. Trained ML algorithms use such *w_i_+* and *w_i_-* and the other features as arguments to predict proRate. Factor SHAP values are calculated based on the aforementioned factor weights, features, and trained models.

To ensure the highest standardization, digitization, and open-access characterization of the evaluation reports, the method only considers projects that ended between 2004 and 2024.

IV. DATA INPUTS

A. PROJECT EVALUATION REPORT CORPUS

In this study, we prioritize factors important for partnership success by analyzing three data inputs: (1) a corpus of project evaluation reports; (2) project ratings assigned by independent evaluation office experts from individual organizations, based on each organization's evaluation guidelines (as explained in Section IV-B); and (3) factors identified in the literature as described in Section IV-C. This analysis focuses on the cases where co-financing equals 1 (coFin = 1) and project rating equals 4 (proRate = 4).

The corpus comprises 1,135 project evaluation documents[5] in PDF format, downloaded from AsDB, IFAD, UNDP, and WB repositories.[6] All these repositories are open, but require some machinery for bulk download. These organizations were chosen with three objectives in mind: (a) to represent the variety of multilateral organizations by development mandate, geographic scope, and voting power mechanism; (b) to maximize information availability, accessibility, and quality; and (c) to contain the number of organizations in the workable resources of a single person PhD project. Document dimensions range from one page to 328 pages and from 47 kB to 17 MB. Further project information[7] is stored in a project metadata repository.

In the absence of official lists of co-financed and non-co-financed project codes, distinguishing between partnerships (coFin = 1) and non-co-financed projects (coFin = 0) requires different approaches depending on the organization. As far as AsDB is concerned, co-financed projects are filtered out by looking for multilateral organization names in the surroundings of the strings 'source of funding / amount', 'cofin', in the project database, the design documents, and the evaluation reports. Concerning IFAD, text mining and regular expressions (regex) identify co-financed projects by looking for multilateral organization names, monetary values, and dollar currency symbols in the context surrounding the phrases "co-financiers (international)" or "project costs and financing in design or evaluation documents. Concerning UNDP, partnered projects were identified by looking for multilateral organization names among the "donors" in the "budget_source" field of UNDP Application Programming Interface (API).[8] The WB has a project API suitable for batch downloads. Text mining and regular expression help identify co-financed projects in design or evaluation documents by searching for three items: (1) the word "financing" or "financed"; (2) the name of a multilateral organization; and,

[5]This number represents the evaluation documents for all co-financed projects closed in the specified time range (see Section III) that met specific criteria and were accessible via the downloading scripts, along with documents for an equivalent number of non-co-financed projects. The documents include evaluation documents of type: "PVR", "PCR", "PPER", "PPA", "PCRD", "PPE", "MTR", "PE", "ICR", "ICRR", "PPAR".

[6]Project evaluation document repositories, accessed on Aug. 22nd, 2024: AsDB: https://www.adb.org/projects; IFAD: https://www.ifad.org/en/programme-and-project-documents; UNDP: https://open.undp.org/projects; WB: https://documents.worldbank.org/en/publication/documents-reports

[7]Extended project name, second co-financier, budget, approval date, sector, sub-sector.

[8]https://api.open.undp.org/api/project_list or https://api.open.undp.org/api/projects

TABLE 2. Outline of the script that takes in input the PDF corpus of project evaluation reports drawn from open repositories of multilateral organizations, and processes them into factor SHAP values and graphs for factor importance ranking.

Script step	Description	Key packages
1	Text conversion from PDF to text (TXT)	Fitz [26]
2	Retention of the sole evaluation documents in English	langdetect
3	Sentence preparation, separation and numbering	WordNetLemmatizer
4	Sentence sentiment analysis and polarity score	VADER, TextBlob, Stanza, re-trained Stanza, Flair
5	Sentence and factor embeddings	SentenceTransformer, TfidfVectorizer
6	Sentence count by factor and polarity with sBERT embeddings (similarity tables)	cosine_similarity
7	Sentence count by factor and polarity score with TF-IDF embeddings (similarity tables)	cosine_similarity, scipy.sparse
8 sBERT	Individual factor weighing by polarity and semantic similarity with sBERT embeddings	Pandas for dataframe operations
8 TF-IDF	Individual factor weighing by polarity and semantic similarity with TF-IDF embeddings	Pandas for dataframe operations
9 LOOCV	Models optimization, training and Leave One Out Cross Validation (LOOCV) evaluation	optuna, LogisticRegression, RandomForest-Classifier, XGBClassifier, MLPClassifier
9 RCV	Models 5-fold Randomized Cross Validation (RCV) evaluation	StratifiedKFold
10	Factor SHAP values and graphs	shap
11	Factor SHAP value confidence intervals	Bootstrap Sampling

(3) a numeric monetary value in the surroundings of the name of the multilateral organization.

Of such a large corpus, cf. Figure 1, only 388 documents met the following requirements: be text and not images; be at least 80% written in English; contain sentences longer than ten words; have an openly accessible rating; refer to projects ending in the specified year range; and, ensure that for each co-financed project (partnership) there is a non-co-financed project. The non-co-financed project subset was built to be as close as possible to a control group with the following priority order: country > sector > year. However, it is impossible[9] to have two projects by the same organization, in the same country, in the same sector, completed in the same year, of which one is co-financed by another multilateral organization and the other is not.

The work considered one evaluation report per project, but it can easily be expanded to pertinent ancillary documents (e.g., press releases) simply by collating them in PDF format. The class imbalance (the number of projects by rating class) is inherent in the industry and is emphasized by the adopted rating re-scaling, discussed in Section IV-B. Future research may include more organizations for projects with high ratings and diversity to compensate for class imbalance or merge the re-scaled proRate = 3 and proRate = 4 into a single class.

B. PROJECT RATINGS

The proRate, used as a categorical feature, is a composed index that summarizes project and partner performance on a gradual four- or six-class scale, as formalized in each organization's evaluation guidelines.[10] Thus, proRate is the outcome for ML to predict for both partnerships and non-co-financed projects. The four-class textual and float number ratings are re-scaled as shown in Table 3. For standardization and reduction of the class number, this study re-scaled the six-class scale into the four-grade scale according to Table 4 and translated textual ratings into integers: highly unsatisfactory = 1; unsatisfactory = 2; satisfactory = 3; and, highly satisfactory = 4.

TABLE 3. Re-classification of float ratings into four textual ratings and proRate integers.

Float ratings range	Textual ratings	Re-classified ratings (proRate)
3.50 - 4.00	highly satisfactory	4
2.50 - 3.49	satisfactory	3
1.50 - 2.49	unsatisfactory	2
1.00 - 1.49	highly unsatisfactory	1

We highlight that rating re-scaling and the thresholding system tend to rarefy the number of projects with proRate = 4.

Matching ratings (proRate) and project codes (proCode), which are unique project identifiers (ID), is not trivial for several reasons: the overall rating score has a different name by organization[11]; rating lists may contain items that differ from projects[12] and are not relevant to our method; proCode may be presented in different formats[13]; and independent evaluation offices may rely on project names or unique evaluation ID that differ from the project ID. The absence of published maps linking project and evaluation IDs requires the creation of name-based semantic maps and cross-checking of year and country information (see Section V-A).

[9] If that were so, it would equate to a duplication, for inefficient use of public spending.

[10] Project ratings webpages, accessed on August 22nd, 2024: AsDB: https://data.adb.org/dataset/success-rates-database; IFAD: https://ioe.ifad.org/en/independent-evaluation-ratings-database; UNDP: For Ratings: https://aida.undp.org/sdg, do "action" / "export all", concatenate SDG by SDG, move into Excel; WB: https:/ieg.worldbankgroup.org/ieg-data-world-bank-project-ratings-and-lessons

[11] AsDB: Latest overall rating; IFAD: Project Overall Achievement; UNDP: quality rating; WB: IEG Outcome Ratings.

[12] For instance, country programme evaluations or portfolio evaluations in UNDP rating data.

[13] E.g., IFAD databases present projects with ten digits in the project database and with four digits in the evaluation database.

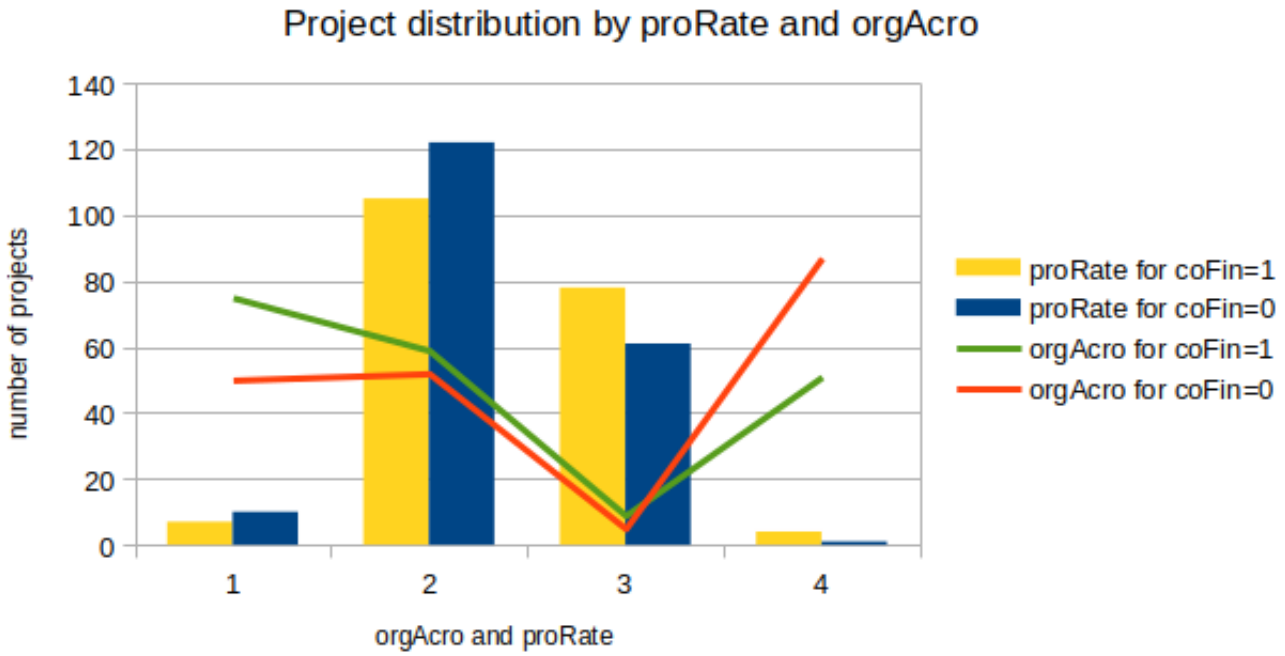


FIGURE 1. Corpus' project distribution by project co-financing status (coFin), overall rating (proRate) and organization (orgAcro). The bars relate to the number of projects by proRate, and the lines relate to the number of projects by orgAcro, from left to right: AsDB = 1; IFAD = 2; UNDP = 3; WB = 4.

These challenges highlight the potential for multilateral organizations to improve their open data practices, adopting standard templates, formats, names, and classifications while also publishing project and evaluation code maps and improving the interoperability of project and evaluation databases.

C. FACTORS THAT INFLUENCE PARTNERSHIP PERFORMANCE

The third data input is an initial list of 753 raw factors[14] drawn from a systematic literature review and a meta-analysis of thematic evaluations of development partnerships as anticipated in Section I-A, which we carried out in the initial part of the work. The initial part of the work [5] provided definitions and theoretical background for partnerships and used NLP techniques, principal component analysis, and K-means clustering for dimensionality reduction, topic modeling, and visualization. The candidate factors concern three key dimensions: (1) operations in the geographical context; (2) governance and management; (3) project quality, and may be aggregated into 10 to 15 clusters. The factor list contains redundancies (duplicates, conceptual similarities, or overlaps), factor nesting, and nonexplicit interrelated effects on partnership performance. It also highlights the limitations of a solely modifier-based semantic analysis to distinguish factor criticality, driving to the more sophisticated method described in this paper.

This factor list allows us to jumpstart the work described in this paper. Future consolidation work shall subject the factors to a thorough ontological and taxonomic review.

V. TOOLKIT

A. CORPUS DOWNLOADING SCRIPTS

Corpus downloading and filename formatting are enabled by 22 scripts, which rely on organizations' APIs when available, such as the WB and UNDP. The scripts, threaded by project country, do the following:

- Filtration of relevant projects by country and project end year with string-matching work done with *jellyfish*'s *Jaro-Winkler Distance* [27] similarity with ratio 0.8;
- Download design and evaluation documents with *Requests* [28];
- Analysis of project HTML page tree to identify the downloading links to the relevant project evaluation or design documents with *BeautifulSoup* [29];
- Classification of the documents by project co-financing status;
- Identification of the co-financiers, looking into document tables, cross-checking, and storing data from web repositories;
- Identification of the project code;
- Elimination of duplicates, retaining only the document in the repository of the main financier, prioritizing documents by exhaustiveness and stability.[15]

[14]Three examples of factors are: "mechanisms for expression of stakeholder viewpoint on the partnership"; "cost-benefit analysis"; and, "harmonization of procurement and disbursement procedures".

[15]Decreasing order of priority: full evaluations by independent evaluation offices, then standard evaluations, project completion digests, mid-term reviews, and, finally, supervision reports.

TABLE 4. Project rating re-scaling from a six- to a four-class scale.

6-grade scale World Bank, IFAD, UNDP (Independent Evaluation Group of the World Bank, 2005; IFAD Revised Evaluation Manual – Part 1, 2022; UNDP Evaluation Guidelines, 2021)			Ratings re-scaling from 6-grade to 4-grade scale, based on thresholds	Ratings integer scores and thresholds (AsDB IED) [UNDP IEO]	4-grade scale Asian Development Bank (Independent Evaluation Department of the Asian Development Bank, 2016)		re-scaled rating
Ratings Integer scores (IFAD) [WB]	Ratings textual scale	Simplified Meaning			Ratings textual scale	Simplified Meaning	
6	Highly Satisfactory	all positive	6 ==> 4	4 (≥ 2.5 / 3 i.e. ≥ 83% i.e. ≥ 4.98 on a 6-grade scale) [≥ 3.5 / 4 i.e. ≥ 88% i.e. ≥ 5.28 on a 6-grade scale]	Highly successful	Exceeds expectation	4
5	Satisfactory	positive with minor shortcomings	5 ==> 3				3
4	Moderately Satisfactory or Mostly Satisfactory	positive with moderate shortcomings	3 and 4 ==> 2	3 (1.75 ≤ x < 2.5 i.e. 58% ≤ x < 83% i.e. 3.48 ≤ x < 4.98 on a 6-grade scale) [2.5 ≤ x < 3.5 i.e. 63% ≤ x < 88% i.e. 3.78 ≤ x< 5.28 on a 6-grade scale]	Successful	Minimal negative impacts	2
3	Moderately Unsatisfactory or Mostly Unsatisfactory	significant shortcomings		2 (0.75 ≤ x < 1.75 i.e. 25% ≤ x < 58% i.e. 1.50 ≤ x < 3.48 on a 6-grade scale) [1.5 ≤ x < 2.5 i.e. 38% ≤ x < 63% i.e. 2.25 ≤ x < 3.78 on a 6-grade scale]	Less than successful	Negative impacts outweighing positive impacts	
2	Unsatisfactory	major shortcomings	1 and 2 ==> 1				1
1	Highly Unsatisfactory	severe shortcomings		1 (x < 0.75 i.e. x < 1.5 on a 6-grade scale) [x < 1.5 i.e. x < 2.25 on a 6-grade scale]	Unsuccessful	Negative impacts substantially outweigh any positive impact	

Commentary to Table 4: Based on AsDB and UNDP ratings integer scores and thresholds, it is possible to average inter-class thresholds. For instance, see the fifth column of the top row of Table 4: according to AsDB IED's four-grade scale, to move from rating 3 to rating 4, the project should have a score not inferior to 2.5 out of 3 (i.e., 83%). On a six-grade scale, 83% of the top rating 6 is 4.98. In UNDP IEO's four-grade scale, moving from rating 3 to rating 4 requires that a project has a rating not inferior to 3.5/4 = 88%. On a six-grade scale, 88% of 6 is 5.28. The average between 4.98 and 5.28 is 5.15. Therefore, only a rating of 6 exceeds 5.15 and may be admitted to class 4 on a four-grade scale. Rating 5 stays below the 5.15 threshold and would be admitted to class 3.

- Renaming of the retained documents as per the work's standard with all parameters necessary for the main script: *couCode*; *proYear*; *docType*; *coFin*; *orgAcro*; *proCode*, and
- Reconstruction of file names from JSON-mapped strings.
- When no direct connection is provided, a project is matched with its independent evaluation by semantic comparison of characteristics.
- Creation and arrangement of corpus directories to systematically store documents downloaded from various organizational websites.

- Control corpus creation, choosing control projects through characteristics matching with corpus' co-financed projects.

B. MAIN SCRIPT

The main script, compiled in Python 3.10.12 and summarized in Table 2, calculates the priority factors by deriving their associated SHAP values from PDF documents. It consists of 11 sequential steps, including corpus preparation, sentiment analysis, semantic similarity analysis, factor double weighing, ML, evaluation, factor SHAP value, and bootstrap sampling confidence interval calculations. The evaluation step assesses the rating classification performance of this method for successful partnerships (coFin = 1, proRate = 4) against the current success rate for co-financed project partnerships (81%) and knowledge partnerships (75%). The SHAP value of a factor measures its contribution to the rating.

The main script is organized as follows. The algorithmic models were selected to ensure the evaluation of a variety of models while minimizing their number according to the following criteria:

- Ensuring at least a full XAI and a full opaque algorithmic chain;
- Including multiple machine learning architectures: linear, tree-based, and neural, and multiple sentiment analysis architectures: lexicon- and rule-based and neural;
- Allowing for CPU calculation to ensure higher availability, affordability, and sustainability for use in low-resource contexts than GPUs; and,
- Ensuring capacity for processing large sentences and institutional jargon.

The script uses the following sentence filtration parameters, enabling sensitivity analyses:

- An 80% threshold for English sentences is imposed to retain the document for further processing.
- Lemmatization is preferred for preserving the nuances of the sentences. Transformer-based sBERT embeddings capture these nuances.
- Stopwords with negative or adversative[16] sense are retained in sentences because they express critical judgments that characterize evaluative documents. This is achieved by removing the negative and adversative stopwords from the stopword dictionary. Tests are performed with and without these stopwords.
- *senLeng*: Two values are tested: senLeng = 10 and senLeng = 20.
- *neuTres* is tested at 0.1 with TF-IDF embeddings and 0.4 and 0.6 with both TF-IDF and sBERT embeddings.
- *simRatio* is tested in the 0.3-to-0.9 interval with 0.1 increases with opaque embeddings and in the 0.1-to-0.5 interval with 0.1 increases with TF-IDF embeddings.
- *orgAcro*: The main financier name is intentionally dropped.

A corpus-level sentence counter *senNGlo* and document-level sentence counter *senNDoc* (Step 3) enable future sentence traceability, improving explainability.

Running all script steps in a notebook from scratch requires 65 hours in 10-factor simulations or 142 hours in 753-factor simulations in a 24-CPU Linux machine. These durations may be reduced to 2 and 40 hours by using scripts instead of notebooks and reusing the document preparation and sentiment analysis output of Steps 1 to 4. Screen sessions simplify the implementation. The maximum RAM consumption is 85 GB in Step 4.

The current version of the proposed script trains predictive models *without* the organization name (orgAcro) and with a combined couCode-proYear context feature to make the model organization-agnostic and avoid inter-organization or inter-country comparisons, opening up to predictions for virtually any organization.

In summary, the script outputs are four ML predictive models (LR, RF, XGB, and NN) in Joblib format trained and evaluated with LOOCV and five-fold RCV methods, factor SHAP values in h5 format, and a set of graphs (summary, importance, force plots, and confusion matrices) by model, proRate, and coFin.

1) ML ALGORITHMS

We now explain the criteria for selecting ML algorithms and setting hyperparameters in the main script.

a: ALGORITHM SELECTION

The four selected algorithms represent four explainability grades from highly explainable LR to opaque NN through RF and XGB–two forest-based models. The Optuna framework was used to define and optimize the model hyperparameters [30], [31] with 30 trials per model to balance the training time and the number of hyperparameters to optimize.

b: LOGISTIC REGRESSIONS SELECTION CRITERIA

LR calculate the probability that a predicted proRate belongs to a particular class (1, 2, 3, or 4) more sensitively than linear regressions due to the S shape of LR, close to but never below zero, and close to but never beyond one [32]. LR are regarded as a highly interpretable and explainable model because they use a linear combination of input features, and their coefficients render the strength and direction of the factor-rating relationship. LR are better suited to binary qualitative responses than linear regressions [32].

[16]Negative and adversative words excluded from stopword removal in all simulations but simulation-05 (sim-41 data): "but"; "M&E"; "should"; "while" ; "whereas"; "more"; "against"; "not"; "cannot"; "can not"; "will not"; "won't"; "would not"; "wouldn't"; "should not"; "shouldn't"; "could not"; "couldn't"; "did not"; "didn't"; "does not"; "doesn't"; "do not"; "don't"; "have not"; "haven't"; "has not"; "hasn't"; "had not"; "hadn't"; "were not"; "weren't"; "was not"; "wasn't"; "are not"; "aren't"; "is not"; "isn't"; "fail to"; "lack of"; "without"; "absence of"; "fall short".

c: LOGISTIC REGRESSIONS HYPERPARAMETER EFFECTS

This study uses a *cross-entropy loss* likelihood function with parameters "*liblinear*" with sparser datasets and "*saga*" in large datasets. Parameter C is highest (77.49) in the 10-parameter Simulation 07, ranging from 0.14 to 2.01 in the other simulations. This high C value increases the risk of overfitting, whereas a lower-value range mitigates it. In Simulation 08, C = 1.1363 imposes a moderate regularization, balancing underfitting and overfitting.

d: RANDOM FOREST SELECTION CRITERIA

This study aims to prepare for future research on the causality of partnership factors. RF selection is based on the work of Athey et al. [16] and Microsoft [33] on forest-based models for causal analysis. RF assign a prediction to the most commonly occurring proRate class by minimizing the *Gini impurity* index [32], the current default criterion in *scikit-learn*'s *randomforestclassifier*. RF work well in classification, are better than individual trees, and have an accuracy comparable to the NB classifier with data sets with many weak inputs [34] as they occur in this work. However, NB classifiers assume factor independence, which is not completely true in real-world partnerships. Decision trees, a tree-based alternative explainable model, tend to overfit in complex models such as the 1,506 factor-based weights of this work, whereas RF "always converge so that overfitting is not a problem" [34].

e: RANDOM FOREST HYPERPARAMETERS EFFECTS

A larger number of trees (*n-estimators*) reduces the variance and increases generalization and computational time. The higher the maximum tree depth (*max_depth*), the higher the model capacity to capture the complexity and the risk of overfitting. A shallow depth balances the complexity of the model and the risk of overfitting. A high minimum number of samples to split an internal node (*min_samples_split*) makes the models less prone to overfitting. A moderate min_samples_split value, such as 9, prevents excessive splitting and model complexity. The *min_samples_leaf* parameter specifies the minimum number of projects in a leaf node, which is the endpoint of the decision tree where a prediction is made. The higher the *min_sample_leaf* value, the lower the overfitting risk; Simulation 06 has a mid value, whereas the other simulations have a very low value. When bootstrap sampling (*bootstrap*) is turned to true, each tree is built from a sub-set of projects with replacement. In a replacement, certain projects may be accounted for more than once in the tree dataset, and, as a consequence, other projects are not, increasing model robustness and generalization. In summary, hyperparameters may be tuned to capture complexity while limiting overfitting or to capture high complexity, risking overfitting.

f: EXTREME GRADIENT BOOSTING SELECTION CRITERIA

XGB is well-suited for sparse datasets and Central Processing Unit (CPU) machines [35], such as those used in this study. Breiman's question whether "gains in accuracy can be gotten by combining random features with boosting" [34] and the comforting results of Shafiezadeh et al. [30] confirmed the choice. XGB performs well with imbalanced and complex datasets with non-linear relationships and sparse data [35], such as our four rating classes, interrelated factors, and sparse weights resulting from TF-IDF embeddings and TextBlob. XGB processes hundred million examples within time- and memory-bound individual CPU computing resources [35]. The model minimizes a regularized-object function composed of a convex loss function, a function that penalizes the complexity of the model, and a final weight smoothing function that mitigates overfitting [35]. The model uses weight shrinkage and feature sub-sampling as additional overfitting-reduction techniques [35]. To optimize memory usage, instead of enumerating all possible tree splits with an exact greedy algorithm, XGB prioritizes features to process by importance following a "*justified weighted quantile sketch for approximate learning*" [35]. XGB inherently re-addresses the missing values of the sparse dataset to the best default values, which it learns directly from the data [35]. Gradient Boosting Machines, an alternative to XGB, use a greedy algorithm that is less optimized for speed [35] than XGB.

g: EXTREME GRADIENT BOOSTING HYPERPARAMETERS EFFECTS

The Optuna-optimized hyperparameters *lambda*, *alpha*, and *eta* indicate very low L2 regularization (Ridge) and L1 regularization (Lasso) and a relatively high learning rate and risk overfitting. For example, values such as = 0.00057, alpha = 5.08×10^{-6}, and *eta* = 0.3986 could lead to overfitting in this noisy dataset.

The *dart* booster applies tree dropouts and mitigates the risk of overfitting. A low *eta* learning rate (for example, in the range of 0.0398-0.1737) fits the data carefully, and a very low *alpha* L1 regularization (e.g., 0.0142) and moderate *lambda* L2 regularization (e.g., 0.0318) reduce overfitting risks. A very high lambda (e.g., 0.9603) with the dart booster reduces complexity and large-weight impacts.

h: NEURAL NETWORKS SELECTION CRITERIA

NN are suitable for multiclass classifications in datasets with hidden patterns between numerous potentially interrelated factors. NN are composed of an input layer, an output layer, and one or more hidden layers (*n_layers*) made of neurons or *perceptrons*. The output is generated by forward propagation, layer by layer, with a non-linear *activation function* that adjusts the neuron weights to minimize the cross-entropy loss [32] used by the scikit-learn *MLPClassifier* and the *softmax* function in the output layer [36]. This dataset has 1508 input units (the 753 *w_i_+* and 753 *w_i_-* factor weights, *coFin*, and *couCode-proYear*) and four output units (proRate classes from 1 to 4). Compared to Support Vector Machines, an accurate alternative, NN maintain multiclass prediction and computational performance in the case of

non-linear factor dependencies or redundancies and a high number of factors compared to the number of records. NN inherently filter out redundant factors without using additional factor dimensionality reduction techniques.

i: NEURAL NETWORKS HYPERPARAMETERS EFFECTS

A high *n_layers* number or neuron number (*n_units_layer*) allows the model to learn complex patterns, but increases overfitting risk and training duration. NN are affected by the vanishing gradient problem, in which neuron weights remain the same after a certain number of training dataset uses (epochs), preventing accuracy gains, and the exploding gradient problem, in which the network becomes unstable, preventing convergence and learning [9]. This study uses three activation functions (shortened to a.f. in this paragraph): the *logistic* a.f., which is sensitive to the vanishing gradient problem, and the *relu* a.f. and *tanh* a.f., which mitigate both problems. NN require optimization parameters to minimize a cost function that includes a performance measure evaluated on the entire training set and depends on a per-example loss [9]. This work utilizes three training optimization algorithms: Stochastic Gradient Descent (*sdg*); Adaptive Moments (*adam*); and, Limited Memory Broyden–Fletcher–Goldfarb–Shanno (*lbfgs*). Higher values of the regularization parameter *alpha* penalize large neuron weights, mitigating the risk of overfitting. The *learning_rate* parameter controls how fast the model learns [9] and, in this study, has the following values: fixed or *constant*; decreasing (*invscaling*); or *adapting* based on the model's performance. In neural networks with two hidden layers (when combined with XAI) or three hidden layers (in opaque configurations), a high number of neurons per layer (e.g., 96, 94) may suffice to capture non-linear interactions among factors. Conversely, a low number of neurons (e.g., 4-17) can lead to underfitting, as the model may not have sufficient capacity to learn complex relationships.

2) RE-TRAINED STANZA

Corpus sentence polarity scores are a fundamental part of the proposed method. The inclusion of sentiment analysis in the method is rooted in the concepts of success and failure of a partnership and qualitative classification of factor criticality in the literature. Sentiment analysis is performed by AsDB [37], UNDP [38], and WB [23], [24]. Polarity scores are assigned by five algorithms, including a specifically re-trained version of *Stanza*, a Convolutional Neural Networks sentiment analysis tool created by the Stanford NLP Group–where NLP stands for Natural Language Processing–and licensed under the Apache License, Version 2.0, which allows adding new sentiment models. Stanza was chosen because its base version was rated better than the other three models against domain expert polarity scores on a sample document.

A 165,516-paragraph training dataset was created by aggregating and exporting the available sentences with polarity scores from UNDP AIDA [38], [39], an AI-powered development cooperation evaluation paragraph analytics repository produced and maintained by UNDP. We trained and tested five Stanza versions with different configurations of UNDP AIDA sentence sentiment labels, achieving an accuracy range of 0.53-0.77. The second version (V2) had three sentiment labels and high accuracy, providing a proportionally larger number of positive or negative polarity sentences (see Table 5). Therefore, V2 was used in this study.

TABLE 5. Evaluation of the Stanza sentiment analysis model after re-training.

Stanza re-trained V2				
[no. of sentences]		predicted polarity score		
		0	1	2
actual	0	1160	416	44
	1	544	3676	705
	2	17	420	1247
Stanza standard version				
[no. of sentences]		predicted polarity score		
		0	1	2
actual	0	958	558	104
	1	713	3716	868
	2	180	814	690

Commentary to Table 5: On the upper part, Stanza re-trained version V2's confusion matrix compared to the polarity scores of a sentence dataset built from UNDP AIDA; the accuracy is 0.74. On the lower part, the Stanza standard version's confusion matrix is based on the same sentence dataset, and the accuracy is 0.57. AIDA is courtesy of UNDP. Stanza is courtesy of Stanford NLP group, Apache license V2. Stanza returns three scores: 0 for negative sentiment; 1 for neutral sentiment; and, 2 for positive sentiment.

VI. FACTOR RANKING

A. SIMULATION OUTLINE

Following Fryer et al. [40], actor prioritization is "*motivated by the notion that a submodel exists that is of higher value than the full model.*" This improves the model's predictive performance and yields a more cost-effective predictor. Factor selection aims to curb appraisal costs while preserving reliability.

The multiclass project rating classification performance of four ML (LR, RF, XGB, and NN) models is evaluated with LOOCV and RCV in seven parameter configurations called *simulations*. The simulations and their parameters are presented in Table 1.

The simulation evaluation is based on the multiclass MCC [41], because of its balance with negative values and class imbalances. MCC gives a high score only if the ML prediction obtains good results in all categories of the confusion matrix (true positives, false negatives, true negatives, and false positives) relative to the size of each of the four subsets identified by the four *proRate* grades. Because this method aims to equal or exceed the current partnership success rate (81% for project co-financing partnerships, 75% for the other partnership types), a combination of algorithms

and parameters is accepted when $MCC \geq 81\%$ in the RCV evaluation. We used Equation (1) for MCC:

$$\text{MCC} = \frac{\sum_{i,j,k,l=1}^{4} \epsilon_{ijkl} C_{ij} C_{kl}}{\sqrt{\sum_{i=1}^{4}\left(\sum_{j=1}^{4} C_{ij}\right)\left(\sum_{k\neq i}^{4}\sum_{l=1}^{4} C_{kl}\right)}} \cdot \frac{1}{\sqrt{\sum_{j=1}^{4}\left(\sum_{i=1}^{4} C_{ij}\right)\left(\sum_{l\neq j}^{4}\sum_{k=1}^{4} C_{kl}\right)}} \quad (1)$$

where:

- C_{ij} is the element of the confusion matrix corresponding to actual class i and predicted class j.
- C_{kl} is the element of the confusion matrix corresponding to actual class k and predicted class l.
- ϵ_{ijkl} is the Levi-Civita symbol, which equals +1 for even permutations, -1 for odd permutations, and 0 if any index is equal.

Other metrics, such as those[17] utilized by UNDP in developing the AIDA sentiment analysis [38], were calculated but are not discussed here. In particular, we calculated Precision, Recall, and F1 Score using the weighted averaging method to account for the highly imbalanced class distribution, ensuring that the overall metric reflects the real-world prevalence of each class.

B. FACTOR SELECTION

Simulations 01 and 04, with the entire factor set, and their derivatives 07 and 08, with the Top-10 critical factor set, explain the factor ranking by criticality.

Simulation 01 (Table 6) is the starting point for subsequent variations because it uses the largest sentence dataset, median neuTres and simRatio parameters, and accurate opaque algorithms. This configuration yields MCC = 0.76 in the RCV with the XGB model. We interpret the lower performance of the other models, including the comparatively good LR performance, as the effect of factor redundancies and interrelations with a relatively limited number (388) of records, highlighting the potential for factor simplification.

The other essential simulation is Simulation 04 (parameters and evaluations in Table 7), which is based on an XAI algorithm chain and the whole factor set. In this simulation, the best-performing explainable ML model, LR, achieved MCC = 0.537, whereas a mixed explainable-opaque chain scored MCC = 0.774 with XGB. Although TF-IDF detects the keywords in the corpus and factor dictionaries and cuts part of the noise due to factor redundancies and interrelations, factor noise may still force the MCC of RF, LR, and NN between 0.527 and 0.566 in RCV.

We now demonstrate how selecting a subset of critical factors maintains or improves ratings-prediction performance while simplifying a partnership's appraisal.

From cooperative game theory, the Shapley values quantify each factor's contribution to an ML model's prediction by considering all factor combinations. In factor selection, the Shapley values rank factors according to their importance [40], accounting for interactions. Shapley values consider how each factor's contribution changes when considered alongside different subsets of other factors, capturing not only the independent impact of a factor but also its effect when interacting with others. This allows for assessing the true importance of each factor in the model's predictions. The SHAP values are calculated in Step 10 of this work with TreeExplainer for XGB and RF and KernelExplainer for the NN and LR models.

[17] Accuracy, prediction, recall, and F1 score.

TABLE 6. Simulation 01.

Simulation 01 conditions: 753 factors; senLeng>10; neuTres: 0.4; simRatio: 0.3; embeddings: sBERT; sentiment analysis: re-trained Stanza.								
Best Logistic Regression Hyperparameters: {'C': 0.218070749279935; 'solver': 'saga'}								
Best Random Forest Hyperparameters: {'n_estimators': 96; 'max_depth': 24; 'min_samples_split': 7; 'min_samples_leaf': 6; 'bootstrap': False}								
Best XGBoost Hyperparameters: {'booster': 'gbtree'; 'lambda': 0.0005723870435249154; 'alpha': 5.083802925413373e-06; 'eta': 0.3986174966081444}								
Best Neural Networks Hyperparameters: {'n_layers': 2; 'n_units_l0': 96; 'n_units_l1': 94; 'activation': 'relu'; 'solver': 'sgd'; 'alpha': 0.030264172310747312; 'learning_rate': 'adaptive'}								
Model	**LR**		**RF**		**XGB**		**NN**	
Evaluation	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**
Accuracy	0,824	0,810	0,725	0,715	0,885	0,882	0,815	0,801
Precision	0,781	0,774	0,690	0,684	0,874	0,878	0,769	0,763
Recall	0,824	0,810	0,725	0,715	0,885	0,882	0,815	0,801
F1 Score	0,794	0,779	0,689	0,673	0,877	0,875	0,787	0,771
MCC	0,653	0,628	0,444	0,419	0,779	0,776	0,633	0,608

Commentary to Table 6: LE stands for LOOCV training and Evaluation, and RE stands for RCV training and Evaluation. Precision, Recall, and F1 Score are based on the weighted average.

TABLE 7. Simulation 04.

Simulation-04: 753 factors; senLeng>10; neuTres=0.1; simRatio=0.1; embeddings: TF-IDF vectorizer; sentiment analysis: TextBlob								
Best Logistic Regression Hyperparameters: {'C': 2.01268022040545; 'solver': 'saga'}								
Best Random Forest Hyperparameters: {'n_estimators': 187; 'max_depth': 32; 'min_samples_split': 13; 'min_samples_leaf': 1; 'bootstrap': True}								
Best XGBoost Hyperparameters: {'booster': 'dart'; 'lambda': 3.215967860855042e-08; 'alpha': 9.394220307682626e-05; 'eta': 0.298959373948566}								
Best Neural Networks Hyperparameters: {'n_layers': 2; 'n_units_l0': 27; 'n_units_l1': 42; 'activation': 'logistic'; 'solver': 'lbfgs'; 'alpha': 0.01623086442185028; 'learning_rate': 'invscaling'}								
Model	**LR**		**RF**		**XGB**		**NN**	
Evaluation	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**
Accuracy	0.793	0.768	0.773	0.756	0.885	0.880	0.804	0.782
Precision	0.747	0.724	0.748	0.743	0.878	0.872	0.783	0.747
Recall	0.793	0.768	0.773	0.756	0.885	0.880	0.804	0.782
F1 Score	0.764	0.740	0.738	0.717	0.878	0.869	0.780	0.758
MCC	0.586	0.537	0.556	0.527	0.781	0.774	0.610	0.566

Commentary to Table 7: LE stands for LOOCV training and Evaluation, and RE stands for RCV training and Evaluation. Precision, Recall, and F1 Score are based on the weighted average.

Figure 2 compares Simulation 01's factor summary plots by decreasing SHAP value and partnership success rate (coFin = 1, proRate decreasing from 4 to 1). The Top-10 factors of the top left sub-figure are the most critical factors for partnership success (coFin=1, proRate=4). The method also identifies the critical failure factors, the top ones in the bottom-right summary plot, for proRate = 1.

The plots provide insights into how various factors contribute to proRate predictions. The higher the factor position on the vertical axis, the higher the factor's mean absolute SHAP value and the factor's importance in the prediction. A red point indicates a high *w_i_+* or *w_i_-* value for the factor in a specific project. In contrast, a blue point indicates a low factor value related to a specific project in the input dataset. The more a point is on the right, the more the feature contributes positively towards the concerned *proRate* prediction. The more a point on the left, the more the factor decreases the predicted *proRate*. The higher the arms of a factor, the more the factor influences the prediction. The combined feature *couCode-proYear* is of prominent importance in each graph, highlighting the importance of the context; however, its value is affected by its current simplistic encoding.

We apply the same process to Simulation 04's dataset[18] to identify the Top-10 critical factors for partnership success by SHAP value based on an XAI algorithmic chain. In Simulation 04, LR is used as the reference model because they perform better than RF. The Top-10 XAI critical factors are shown in the SHAP summary plot in Figure 3.

The upper part of Table 8 reports the top critical factors for partnership success as calculated by the opaque algorithmic chain (sBERT, Stanza re-trained, and XGB) on the Simulation 01 data. The lower part of the table lists the ten critical factors identified by the explainable chain (Simulation 04) for comparison. This explains how to distinguish the most important factors, which answers research question **RQ1**.

The incomplete convergence of the critical factors in Simulations 01 and 04 (Table 8) is due to differences in the input datasets and algorithms. More proRate = 4 projects or merging proRate = 4 and proRate = 3 rating classes would increase convergence.

The derivative simulations –Simulation 07 from 01 and Simulation 08 from 04– measure the rating prediction performance of the Top-10 critical factor subsets compared to the whole factor set at coFin = 1 (partnership) and proRate = 4 (success). With 10 factors (Figure 2a and Table 8a), the predictions of Simulation 07 (Table 9) are very close to the predictions of Simulation 01, which uses 753 factors. XGB MCC LOOCV is 0.780 vs. 0.798, RCV is 0.769 vs. 0.785, in a considerably shorter time for computation and information-gathering. Using the 10 fully "XAI" critical factors, as in Simulation 08, achieves an MCC of 75% with RF in LOOCV and RCV within a short execution time (see Figure 3 and Table 10). This is a notable result in two respects: it reaches the current 75% success rate of knowledge partnerships, not far from the current 81% of co-financing partnerships; and it is obtained with explainable algorithms in spite the redundancies and interrelations in the current list of factors, which may be improved by refining factor formulation. Simulations 07 and 08 demonstrate that it is possible to reduce the appraisal time with limited prediction performance loss (Simulation 07) or gains (Simulation 08) by moving to 10 factors from 753 factors. This answers research question **RQ3**.

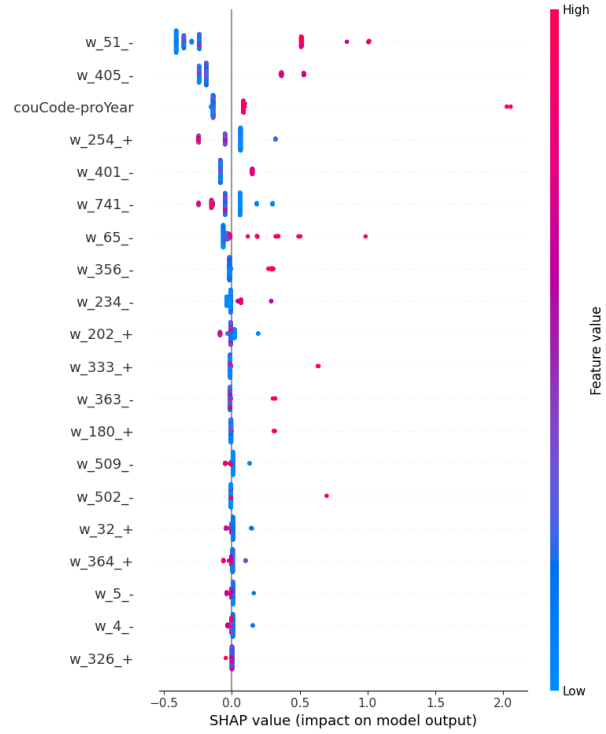
The remaining simulations are sensitivity analyses of the effects of the method parameters. Simulations 02 and 03 show that seeking higher sentence relevance to the factors by increasing the value of the parameters senLeng, neuTres, and simRatio makes the ML dataset sparser, negatively affecting ML's prediction quality or feasibility.[19] The different behavior of the sentiment analysis models explains part of dataset rarefaction (Figure 4): Stanza and Flair tend to concentrate on high polarity scores, TextBlob and VADER on central values, requiring a neuTres threshold sensitivity analysis. TF-IDF causes further data rarefaction. The reasons for this behavior may reside in the TF-IDF vectorizer method. The TF-IDF vectorizer (*tfidf_vectorizer.fit_transform*) requires combining the corpus sentence dictionary and the factor dictionary for later feature dimension consistency, resulting in a sparser one-hot matrix than the sole corpus dictionary. Word frequency and weight may also be lower in wider dictionaries, diluting a word's relevance and lowering similarity scores. The class imbalance of the dataset,[20] which is inherent in the domain, further contributes to the rarefaction of the data, especially in the success and failure classes (proRate = 4, proRate = 1). A correlation analysis between the number of sentences and the MCC LOOCV and RCV evaluation score (not reported here) shows that the predictions of XGB under sparse conditions are the most robust as if XGB's search for weak models compensated for data scarcity. In contrast, RF's are the most sensitive and tend to overfit. Low values of senLeng, neuTres, and simRatio in XAI algorithm chains, such as in Simulation 04,[21] allow sufficient data to be fed to the ML models. Another countermeasure is to reduce the noise by eliminating all zero-value factors from the dataset. However, these factors do not contribute to the final result.

[18] Settings: TF-IDF embeddings; TextBlob sentiment analysis; neuTres = simRatio = 0.1.

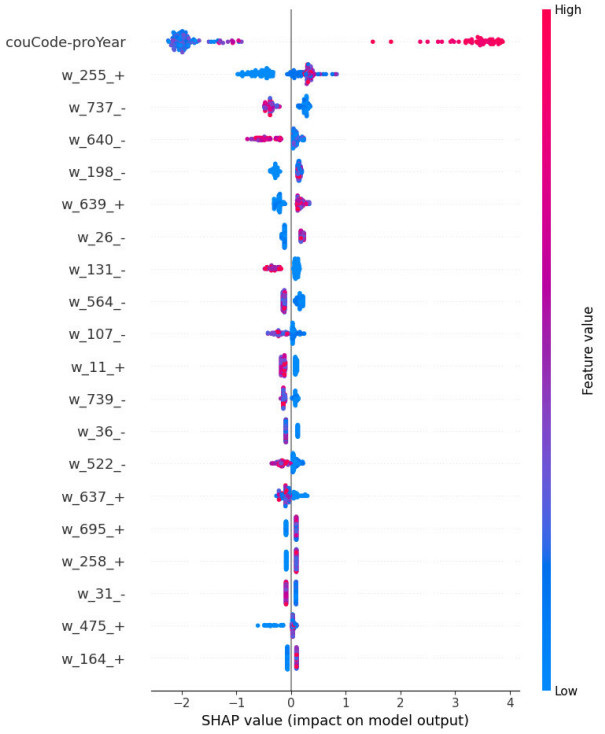
[19] This is the reason why Simulation 04 has a much lower neuTres = simRatio = 0.1 value than Simulation 01, 02, and 03.

[20] Project distribution by individual proRate class: 3% of corpus projects have proRate = 1 (failure); 65% have proRate = 2; 30% have proRate = 3; and, 1% have class 4 (success).

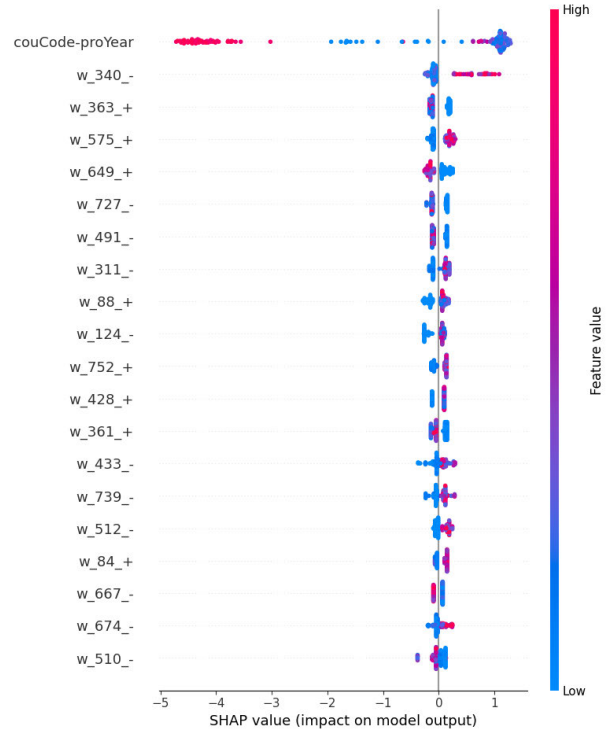
[21] senLeng=10, neuTres=simRatio=0.1.



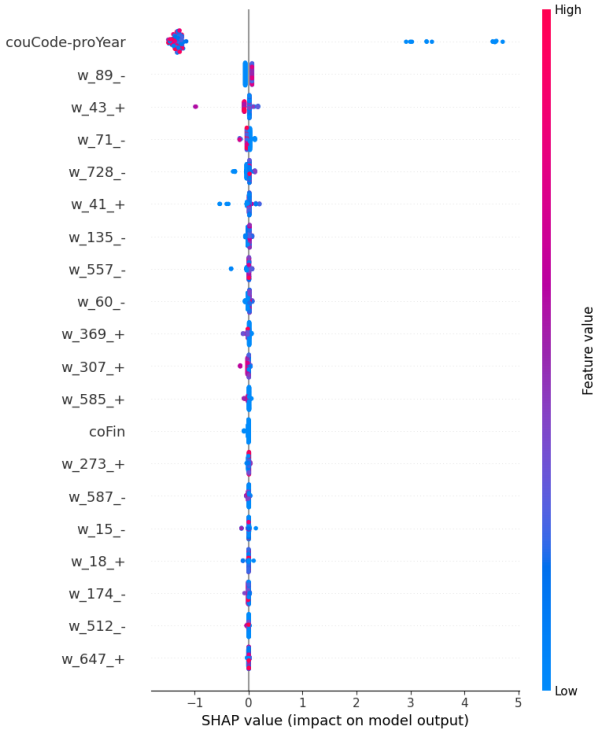
(a) proRate=4



(b) proRate=3



(c) proRate=2



(d) proRate=1

FIGURE 2. **Four-summary-plot comparison (Simulation 01 factors). The plots represent the SHAP value summaries for each class of the proRate variable, from top left to bottom down proRate = 4, 3, 2, 1.**

TABLE 8. Top-10 factors contributing to partnership success (*coFin* = 1, *proRate* = 4).

(a) Simulation 01, opaque algorithm chain: sBERT; re-trained Stanza; XGB

factor	reference
gender-sensitive equity indicators measuring reach and benefit for disadvantaged groups (PPP)	martens-2007
mutual benefit	martens-2007
couCode-proYear	proEval
proportionate liability of all partners in the achievement of success	AfDB-2019
subsidiarity	WB-2011
mutual benefit or mutual resource allocation	mcnamara-2020
long and established history of collaboration	IMF-IEO-2017
complementarity to the activities of donor partners	WB-2011
public disclosure of formalized partnership agreements	WB-2011
inadequacy of resources vis-à-vis the objectives	WB-2011

(b) Simulation 04, XAI algorithm chain: TF-IDF; TextBlob; LR

factor	reference
couCode-proYear	proEval
accessible complaints procedure	martens-2007
mutual benefit	martens-2007
analysis of failed partnership history	ADB-IED-2016
disbursement arrangement harmonization	ADB-IED-2016
focus on long-term capacity building	AfDB-2019
social capital	AfDB-2019
trust building and capacity development	liebenthal-2018
poor leadership	sikandar-2019
parallel approval processes	AfDB-2019

Commentary to Table 8: The factors are ranked by decreasing SHAP value, with their bibliographic references. The *couCode-proYear* is the categorical version of the geographical and chronological context feature taken from the project evaluation document (proEval) names. The upper table (a) shows the critical factors calculated with opaque algorithms in simulation 01. The lower table (b) shows the critical factors calculated with XAI algorithms in simulation 04.

Simulation 05 removes negative and adversative stopwords[22] in corpus sentence preparation, causing a polarity transfer to neutral or positive scores. The effect of the transfer on prediction quality remains unclear due to the sparsity of the ML input dataset, but it deserves further investigation with more inclusive sentence filtration parameters. This study retained negative and adversative stopwords in sentences in subsequent simulations justified by the critical and lexically balanced nature of the evaluation reports jointly prepared by groups of sector experts.

Simulation 06 describes a simplified factor selection method based on the Pareto principle and Zipfian distribution [42]. The simplification consists of ranking factors in descending order according to the number of sentences that are semantically similar to each factor and that possess positive or negative polarity. Although Simulation 06 is simpler and scores a high MCC of 0.788 in RCV, we think it is easier for an organization to collect information on ten factors instead of 24 accepting a slightly lower performance (MCC = 0.749), as in ML- and SHAP-based Simulation 08.

The simulations consistently indicate that ML predictions of partnership performance benefit from an increased number

[22]Negative and adversative words excluded from stopword removal in all simulations but simulation-05 (sim-41 data): "but"; "M&E"; "should"; "while"; "whereas"; "more"; "against"; "not"; "cannot"; "can not"; "will not"; "won't"; "would not"; "wouldn't"; "should not"; "shouldn't"; "could not"; "couldn't" ; "did not"; "didn't"; "does not"; "doesn't" ; "do not"; "don't"; "have not"; "haven't"; "has not" ; "hasn't"; "had not"; "hadn't"; "were not"; "weren't"; "was not"; "wasn't"; "are not"; "aren't"; "is not"; "isn't"; "fail to"; "lack of"; "without"; "absence of"; "fall short."

TABLE 9. Simulation 07.

Simulation 07 conditions: 10 factors based on Simulation 01 SHAP values; senLeng>10; neuTres: 0.4; simRatio: 0.3; embeddings: sBERT; sentiment analysis: Stanza re-trained and re-scaled.								
Best Logistic Regression Hyperparameters: {'C': 77.48595311420625; 'solver': 'saga'}								
Best Random Forest Hyperparameters: {'n_estimators': 160; 'max_depth': 26; 'min_samples_split': 4; 'min_samples_leaf': 1; 'bootstrap': False}								
Best XGBoost Hyperparameters: {'booster': 'dart'; 'lambda': 0.03181540126968269; 'alpha': 7.017458758064946e-06; 'eta': 0.039751009812817 25}								
Best Neural Networks Hyperparameters: {'n_layers': 3; 'n_units_l0': 76; 'n_units_l1': 4; 'n_units_l2': 17; 'activation': 'tanh'; 'solver': 'lbfgs'; 'alpha': 1.6438927561580925e-07; 'learning_rate': 'constant'}								
Model	**LR**		**RF**		**XGB**		**NN**	
Evaluation	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**
Accuracy	0.868	0.866	0.880	0.871	0.885	0.877	0.877	0.880
Precision	0.863	0.857	0.869	0.844	0.876	0.870	0.875	0.889
Recall	0.868	0.866	0.880	0.871	0.885	0.877	0.877	0.880
F1 Score	0.863	0.855	0.867	0.854	0.878	0.869	0.875	0.879
MCC	0.745	0.742	0.769	0.751	0.780	0.769	0.764	0.780

Commentary to Table 9: LE stands for LOOCV training and Evaluation, and RE stands for RCV training and Evaluation. Precision, Recall, and F1 Score are based on the weighted average.

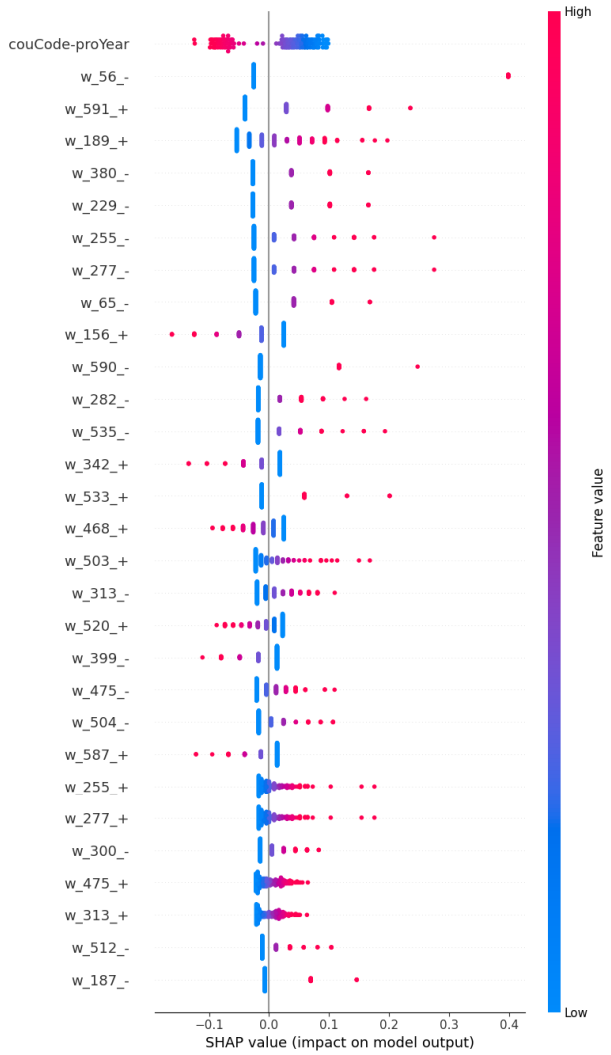


FIGURE 3. SHAP summary plot of Simulation 04 top 30 factors (*coFin* =1, *proRate* = 4, LR). The Top-10 factors are the critical factors for partnership success calculated with an XAI algorithmic chain. Simulation 08 evaluates the proRate prediction performance with such 10 critical factors.

of records per factor, which can be achieved by expanding the corpus and streamlining the factors. Focusing on sentence relevance provides limited benefits and may even be counter-productive.

Table 11 and Figure 5 summarize the evaluation of the predictive performance by simulation.

C. QUALITY OF FACTOR SELECTION

The confidence intervals represent the range within which the true SHAP values are expected to fall, with a confidence level of 95% based on bootstrap sampling. Bootstrap sampling involves selecting projects with replacement. This allows the same project to be chosen multiple times, whereas other projects may not be selected in a given sample. The confidence intervals measure the quality of factor selection by quantifying the uncertainty associated with the estimated

TABLE 10. Simulation 08.

Simulation 08 conditions: 10 factors based on Simulation 04 SHAP values; senLeng>10; neuTres: 0.1; simRatio: 0.1; embeddings: TF-IDF; sentiment analysis: TextBlob.								
Best Logistic Regression Hyperparameters: {'C': 1.13633451952584; 'solver': 'saga'}								
Best Random Forest Hyperparameters: {'n_estimators': 189; 'max_depth': 13; 'min_samples_split': 9; 'min_samples_leaf': 5; 'bootstrap': True}								
Best XGBoost Hyperparameters: {'booster': 'dart'; 'lambda': 0.9602723083082992; 'alpha': 0.014234132378351173; 'eta': 0.17371696369871706}								
Best Neural Network Hyperparameters: {'n_layers': 2; 'n_units_l0': 13; 'n_units_l1': 50; 'activation': 'tanh'; 'solver': 'adam'; 'alpha': 0.0005384201174698013; 'learning_rate': 'constant'}								
Model	**LR**		**RF**		**XGB**		**NN**	
Evaluation	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**
Accuracy	0.866	0.854	0.871	0.868	0.913	0.908	0.874	0.869
Precision	0.820	0.812	0.827	0.826	0.917	0.904	0.873	0.870
Recall	0.866	0.854	0.871	0.868	0.913	0.908	0.874	0.869
F1 Score	0.837	0.827	0.842	0.839	0.909	0.902	0.870	0.863
MCC	0.741	0.717	0.753	0.749	0.835	0.828	0.758	0.751

Commentary to Table 10: Simulation 08 evaluates the rating prediction performance for partnership success with 10 critical factors calculated with explainable algorithms. LE stands for LOOCV training and evaluation, and RE stands for RCV training and Evaluation. Precision, Recall, and F1 Score are based on the weighted average. LE stands for LOOCV training and Evaluation, and RE stands for RCV training and Evaluation. Precision, Recall, and F1 Score are based on the weighted average.

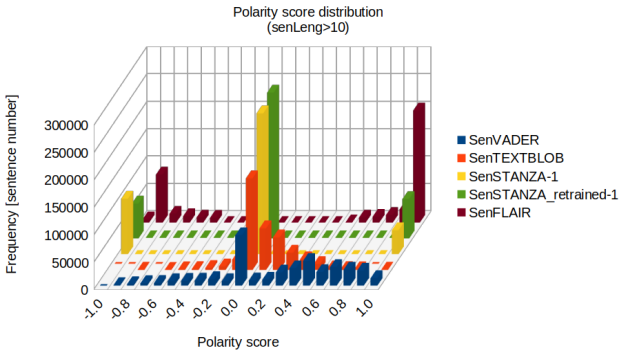


FIGURE 4. Number of sentences by polarity score and sentiment analysis model.

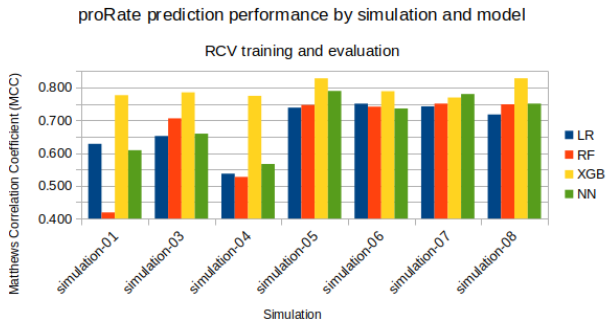
SHAP values, reflecting the variability across multiple samples. The narrower the confidence interval, the higher the confidence in the SHAP value or factor importance. In Figures 6 and 7, confidence intervals are applied to Simulation 01's most important factors, as measured by the XGB model, and Simulation 04's most important factors, as measured by LR.

We define the Relative Confidence Interval (RCI) as the ratio between the confidence interval width and the mean SHAP value, aiming for a normalized comparison of uncertainty across different factors. RCI is calculated

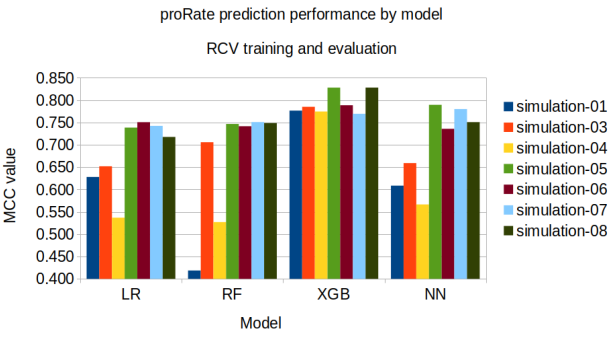
TABLE 11. Summary of MCC evaluations by simulation (accuracy scores are higher than MCC's in all simulations).

Simulation	Data file	Model and evaluation type							
		LR		RF		XGB		NN	
		LE	RE	LE	RE	LE	RE	LE	RE
simulation-01	sim-21	0.653	0.628	0.444	0.419	0.779	0.776	0.633	0.608
simulation-03	sim-01	0.672	0.652	0.724	0.706	0.798	0.785	0.680	0.659
simulation-04	sim-42	0.586	0.537	0.556	0.527	0.781	0.774	0.610	0.566
simulation-05	sim-41	0.755	0.738	0.747	0.746	0.820	0.828	0.778	0.789
simulation-06	sim-37	0.756	0.750	0.740	0.741	0.091	0.788	0.756	0.736
simulation-07	sim-43	0.745	0.742	0.769	0.751	0.780	0.769	0.764	0.780
simulation-08	sim-45	0.741	0.717	0.753	0.749	0.835	0.828	0.758	0.751

Commentary to Table 11: LE means LOOCV Evaluation, RE means RCV Evaluation. Simulation 02 was not evaluated for lack of data. Comparing Simulation 01's XGB RE (MCC=0.628) with the complete factor set to Simulation 07's NN RE (MCC=0.780) with ten critical factors using XAI algorithms demonstrates that using the critical factor subset leads to better and faster performance. This finding is further supported by comparing Simulation 04's LR RE (MCC=0.537) to Simulation 08's RF RE (MCC=0.749) with opaque algorithms.



(a) by simulation



(b) by model

FIGURE 5. MCC RCV evaluation scores by simulation and model.

according to Equation (2):

$$\text{RCI} = \frac{(\text{UCL} - \text{LCL})}{\text{Mean SHAP Value}} \tag{2}$$

where:

- UCL is the Upper Confidence Limit;
- LCL is the Lower Confidence Limit.

In both Simulations 01 and 04, there is a negative correlation between the number of sentences[23] fed to the method and the RCI standard error, variance, and standard deviation, as shown in Table 12). The higher the standard error, variance, and standard deviation, the more spread the RCI, and the higher the SHAP value uncertainty. Because the factors in proRate = 4 (partnership success) have the highest RCI and the lowest number of projects and corpus sentences compared to other rating classes, we conclude that a low number of input sentences causes the uncertainty around SHAP values to vary more widely across factors, lowering the quality of factor selection. This is particularly sensitive for our most relevant rating classes (proRate = 4, proRate = 1).

We conclude that the more data (input sentences) there are, the better the factor selection. Future research may adopt three approaches to decrease the uncertainty of the SHAP value of factors: the inclusion of additional projects in the corpus, a less severe reclassification of project rating for proRate = 4 (Table 4), and a more sophisticated analysis of the critical factors. This analysis may separate the factors important for successful partnerships (proRate = 4, coFin = 1) from those important with proRate $\neq$ 4 and coFin = 1, and coFin = 0.

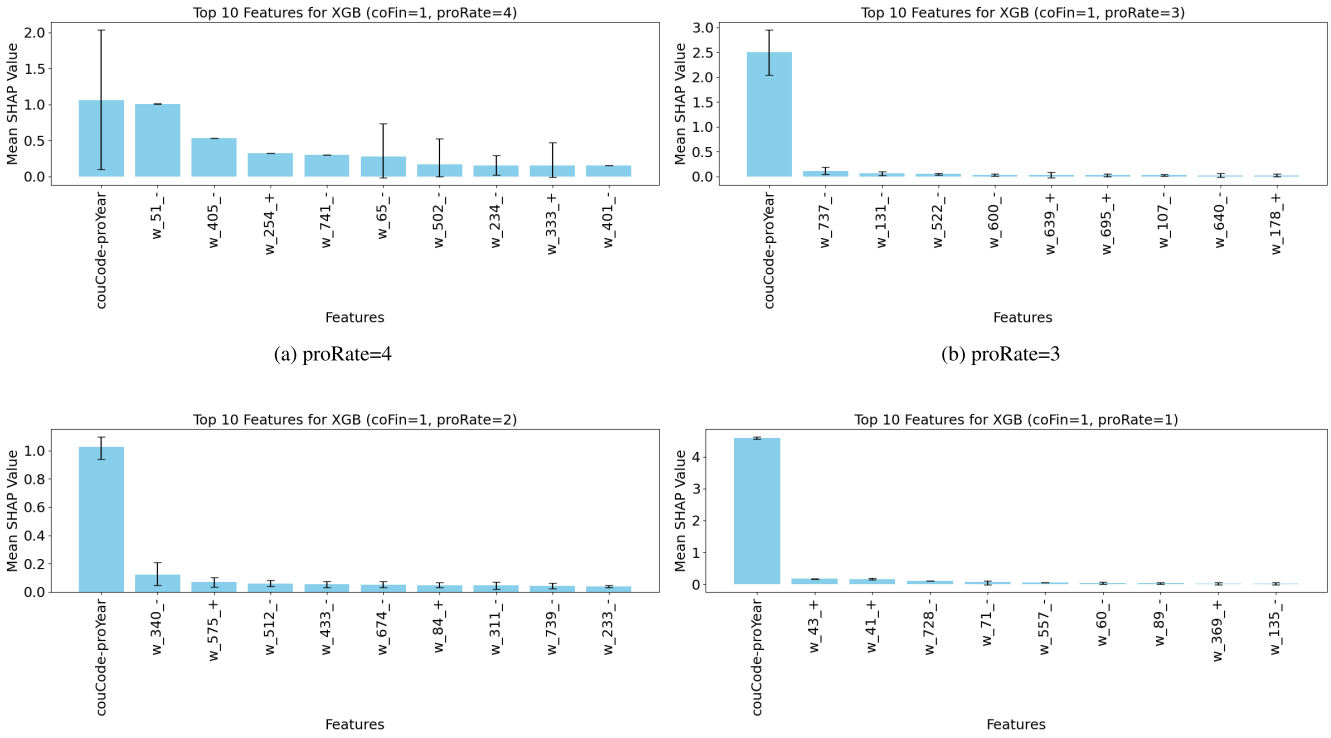
Confidence intervals and RCI measure factor selection uncertainty, answering Research Question **RQ2**.

D. CROSS-CUTTING CONSIDERATIONS

Overfitting risk arises when there is a large difference between the LOOCV and RCV evaluations. We measure the Overfitting Risk Severity (ORS) with Equation (3):

$$\text{ORS} = (\text{LOOCV} - \text{RCV})/\text{RCV} \tag{3}$$

[23]The (a) number of projects, (b) number of rated-project sentences, and (c) number of sentence similarities are substantially interchangeable in the correlations mentioned above since these numbers correlate to one each other at 0.990 or more: corr((a),(b)) = 0.990; corr((b),(c)) = 0.999; corr((a),(c)) = 0.996.



(a) proRate=4 (b) proRate=3

(c) proRate=2 (d) proRate=1

FIGURE 6. Confidence intervals of highest 10 SHAP values for co-financed projects (partnerships) by factor and proRate in Simulation 01 conditions with XGB model and coFin=1. The relevance of and the concerns on the combined geographical and chronological context feature couCode-proYear are discussed in SectionVI-D.

Table 13 shows that ORS is especially high in Simulation 04, suggesting that the model may have memorized the training data. This is reinforced by Kohavi's [43] finding that decreasing k in k-fold RCV implies increasing variance, especially with small sample sizes, and that such variance causes instability. From the simulations, we learned that using the critical factor dataset (Simulations 07 and 08) leads to a lower LOOCV-RCV distance and overfit risk than using the entire factor dataset.

The *contextualization* feature (couCode-proYear) deserves further investigation. Interviews[24] confirmed the critical importance of the context. Consistently, this study found that the contextualization feature couCode-proYear dominates in largely populated rating categories (proRate = 2, proRate = 3). However, the categorical encoding that we utilized was excessively simplistic.[25] Linear models and Neural Networks may attempt to learn linear relationships from normalized categorical feature values despite no meaningful numeric relationship in purely categorical contexts. Tree-based models may try splits based on the numeric order of the normalized categories, negatively affecting meaningfulness and prediction performance. Future research should split couCode-proYear into one or more relevant sub-features, utilizing real values such as UNDP's Human Development Index [44], "*created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country*",[26] or WB's Country Policy and Institutional Assessment values.[27]

The coexistence of opposite-polarity weights in the success (or failure) factor representation, *w_i_+* and *w_i_-*, both valid and sometimes of the same magnitude, may be assessed from a *paraconsistency* perspective to determine whether they are inherent project characteristics or whether they may be substituted by a single value such as Δ w_i = [(w_i_+) - (w_i_-)]. This coexistence suggests that deployed ML models could detect overly optimistic or pessimistic ex-ante partnership design reports and recommend analytical insights for untrustworthy reports. This would further the quantitative use of unstructured data sources.

E. HUMAN-CENTRIC INTERPREPTATION

This method aims to improve partnership success rates and mitigate reputational risks by automatically assessing

[24] Author's interviews to the AsDB IED in 2022, and to the IADB in 2023.

[25] This work identified and assigned 258 unique couCode-proYear values and max-min normalized such values before using them in ML training. 258 couCode-proYear values are numerous compared to the 388 available projects, risking high cardinality.

[26] Yearly dataset published at: https://hdr.undp.org/sites/default/files/2023-24_HDR/HDR23-24_Statistical_Annex_HDI_Table.xlsx

[27] Ref.: https://databank.worldbank.org/source/country-policy-and-institutional-assessment

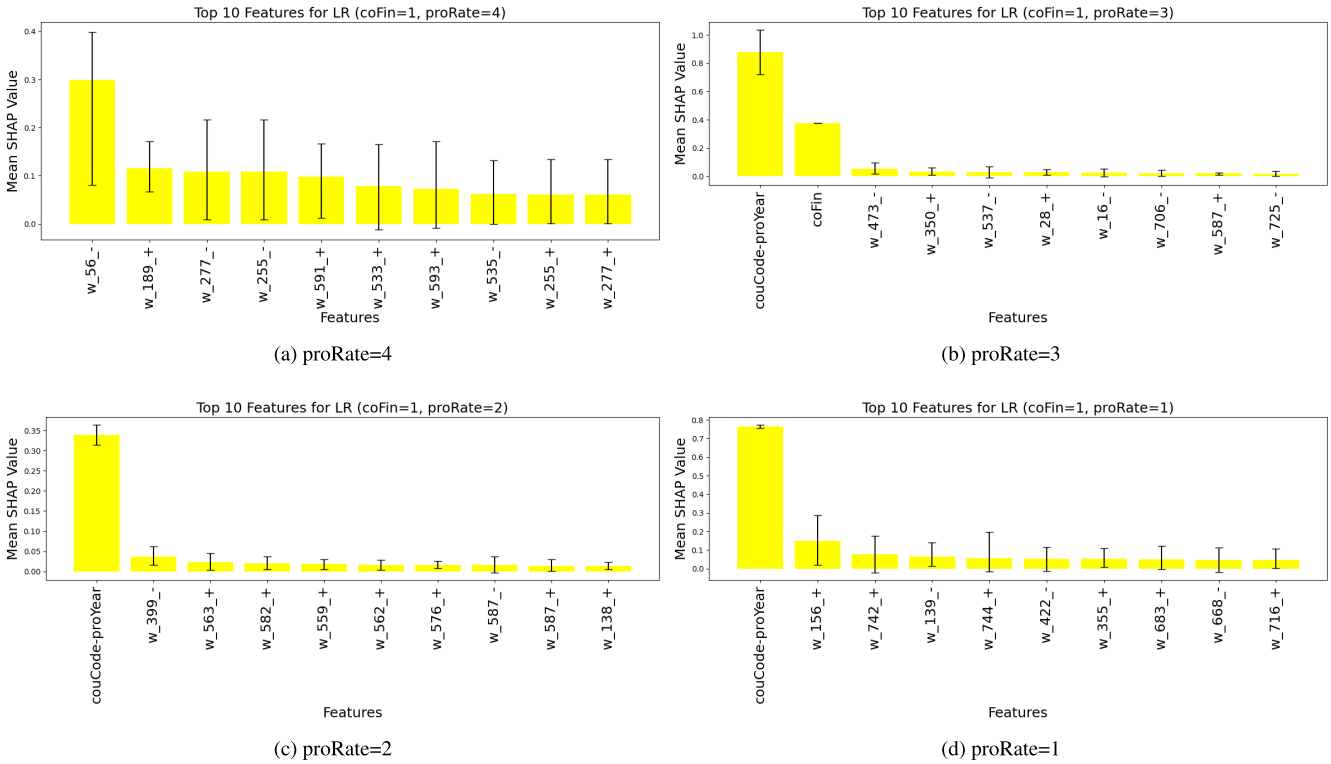


FIGURE 7. Confidence intervals of highest 10 SHAP values for co-financed projects (partnerships) by factor and proRate in Simulation 04 conditions with LR model and coFin=1. On couCode-proYear, see SectionVI-D.

TABLE 12. Statistics on the Relative Confidence Intervals (RCI).

	Re-classified project rating (proRate)				Correlation
	1	2	3	4	
Number of sentences	7,403	117,396	61,954	3,348	
Number of projects	17	227	139	5	
Simulation 01 (opaque algorithmic chain with XGB)					
RCI mean	89.6%	80.2%	162.1%	125.7%	-0.20
RCI standard Error	0.286	0.104	0.285	0.443	-0.88
RCI variance	0.816	0.107	0.813	1.959	-0.81
RCI standard Deviation	0.903	0.327	0.902	1.400	-0.88
Simulation 04 (XAI algorithmic chain with LR)					
RCI mean	218.61%	147.01%	148.77%	187.10%	-0.84
RCI standard error	0.296	0.196	0.278	0.167	-0.15
RCI variance	0.878	0.385	0.773	0.278	-0.20
RCI standard deviation	0.937	0.620	0.879	0.527	-0.15

Commentary to Table 12: The table presents Simulation 01 and Simulation 04 ten most important factors by proRate, and correlations with the number of sentences in rated projects. The individual sentences of this table are counted after sentiment analysis (script's step 4) and, therefore, are counted only once. Only the positive SHAP values are considered; negative SHAP values are not analyzed because we are interested in the factors that most contribute to the concerned proRate.

proposals from project documents from multiple multilateral organizations. With an explainable artificial intelligence (XAI) approach, officers can quickly understand why certain factors, such as country context or disbursement arrangements, are deemed critical for success and, in the future, identify high-potential partnerships. The workflow involves (1) collecting project evaluation or partnership design reports, (2) applying the set of factors and the AI script to prioritize influential factors or generate partnership ratings, and (3) using XAI tools to explain these results. For example, an officer may discover that the low rating of a proposal is due to insufficient details on staff capacity or an unclear result framework and intervene on such factors. Multilateral organizations can help address current model limitations by providing project data in bulk (indicating co-financed vs. non-co-financed) and supporting further research on partnership factor definitions.

TABLE 13. The severity of overfitting risk (ORS) is measured as the relative distance between LOOCV and RCV evaluations by simulation as per Equation (3). The higher the positive relative distance, the higher the overfitting risk.

ORS	LR	RF	XGB	NN	Algorithm chain
simulation-01	4%	6%	0%	4%	opaque
simulation-03	3%	3%	2%	3%	opaque
simulation-04	9%	6%	1%	8%	XAI
simulation-05	2%	0%	-1%	-1%	XAI
simulation-06	1%	0%	0%	3%	opaque
simulation-07	0%	2%	1%	-2%	opaque
simulation-08	3%	1%	1%	1%	XAI

TABLE 14. Potential application domains.

Industry	Input 1	Input2	Input3
Partnerships	Factors	Project evaluation reports	Project Ratings
Research	Evaluation Criteria	Research Proposals	Proposal Scores
Finance	Financial indicators including qualitative ones	Narrative Financial Statement Report	Corporate Ratings

VII. LIMITATIONS AND PRACTICAL APPLICATIONS

The study has limitations with respect to both the validity of current critical factors and the overall usability of the method.

First, the premise that project co-financing represents all partnership types and that MCC can adequately measure performance against current practice success rates are strong assumptions in the sector. As anticipated in sections I-A, IV-C, V-B1, VI-B, the formulation of factors is affected by redundancy, hidden interrelations, double negations, and nesting, which can decrease the accuracy of the classification. Furthermore, the imbalance and scarcity of data in key classes–caused by the inherently limited number of projects, the absence of a specific co-financing flag in project repositories, and our selection criteria–increase the risk of sparsity and overfitting. Although the contextualization feature "couCode-proCode" is critical, it is only coarsely defined as a categorical variable. Together, these issues help explain the gap between the XAI MCC (0.749) and the current success rate (81%) of partnerships, the partial convergence of factor rankings between the XAI and the opaque algorithm chains, and the broad confidence intervals.

Second, applying a chain of XAI algorithms with the current formulation of factors can inadvertently increase partner failures and costs. Partnership performance prediction based on ex-ante documents has not been tried. Although the factors include behavioral aspects (especially related to management attitudes and staff relationships and motivation), a behavioral analysis exceeded the scope of the study. Moreover, the current script is not user-friendly and does not incorporate recommender systems or newer developments such as generative AI technologies.

These limitations can be mitigated by improving the formulation of factors and the contextualization feature and expanding the dataset for better class balance. Evaluating prediction quality based on ex-ante documents and the ML models, which are trained on project evaluation evidence, and further coding can lead to more usable software.

The method presented in this paper assumes that partnerships among multilateral organizations be representative of a variety of organizational domains. Our calculations leverage documents from (four) institutions that span three distinct categories: international financial institutions; multilateral development banks; and UN agencies. We selected those institutions for two fundamental reasons: their intrinsic diversity, and their holding by far the largest publicly available repositories of project documents and ratings. This choice afforded us the ability to generate organization-agnostic factor lists and a vastly diverse training base for machine learning models, from a comparatively small set of considered institutions. We consequently contend that the outcomes we obtained for factor lists and trained models may be applied to other types of organization, including those that do not assign project ratings. This arguably vouches for the broad applicability of our proposed method.

Although the immediate practical utility of our method is limited, once mitigation actions are implemented, prioritized factors can guide and focus multilevel stakeholder discussions on partnerships. The method can shorten appraisal and due diligence processes and enable studies on the effects of project co-financing. Finally, the method can lead to better partnerships, improving the reputation of partners, impact on beneficiaries, and efficiency of public expenditure.

VIII. CONCLUSION

One in five strategic partnerships between multilateral development organizations is currently unsuccessful, but it is costly and less impactful. A list of critical success factors for partnership performance is quantitatively identified and discussed to facilitate multilateral organizations in selecting their future partnerships. The method uses an evaluation report corpus, project ratings, a long list of factors, state-of-the-art Natural Language Processing techniques, and an innovative factor weighting model to identify the Top-10 critical success factors. Using the Top-10 critical success factors, the method's predictions equal the current knowledge-partnership success rate (75%) with explainable algorithms (XAI) and exceed the project co-financing partnership success rate (81%) with a mix of explainable and opaque algorithms. This yields considerable savings in appraisal time and more opportunities for focused discussion. Critical failure factors are also identified using the same method, although they are not discussed in this paper. The quality of the selected factors is measured by relative confidence intervals, which are significant due to current factor redundancies and interrelations. These findings answer the three research questions that prompted this work.

Simulations show that excessive filtering of corpus sentences by relevance can reduce the prediction accuracy of machine learning models and the reliability of factor selection. Instead, it is more effective to simplify factor formulation by using canonization, ontology, and taxonomy. Furthermore, refining the hyperparameters of the XAI algorithms further enhances the performance of the model.

Using *w_i_+* and *w_i_-* positive and negative weights predisposes the method to process ex-ante evaluations.

Future studies that leverage these results may involve various approaches. They may procure more evaluation reports, especially from highly successful or unsuccessful rated partnerships, focus on improving the geographical-chronological context feature *couCode-proYear*, deploy the models, open them up to learn from further evaluations, investigate underfitting or overfitting risks and parameter selection, sentence traceability, effects of exclusion of negative and adversative words from the stopword list, automatic factor detection with part-of-speech or named entity recognition techniques, use of idioms other than English, and causality analysis.

Future research might consider applying this method to Research and Innovation, particularly for evaluating research proposals, and Finance, where it may help predict corporate ratings based on narrative financial statement reports, as shown in Table 14.

APPENDIX
GLOSSARY

coFin is the project co-financing status and takes the value of coFin = 1 for co-financed projects, that is, partnerships, and coFin = 0 for non-co-financed projects. The *coFin* is the treatment feature that distinguishes partnerships according to the adopted definition and assumptions.

couCode is the ISO 3166-1 A-3 three-digit country code string of the International Standard Organization that provides the geographic context at the national level.

couCode-proYear is a contextual categorical covariate that combines couCode and proYear because each country has a peculiar context that may vary by year. In our method, *couCode-proYear* substitutes for the individual couCode and proYear features, which are dropped.

docType is a string indicating the type of evaluation report or design report document, such as PCR for the Project Completion Report or PPER for the Project Performance Evaluation Report.

Opaque, in this paper, is a sequence of algorithms with limited interpretability and higher capabilities in modeling non-linear relationships in data and in capturing relationships between factors: sBERT [45] for embeddings; Stanza [46]; Stanza Re-Trained (see Section V-B2) or Flair [47] for sentiment analysis; and, XGB [48] or NN [49] for predictions. All such models are of the neural type except XGB, which is a boosted ensemble tree-based model. XGB gradient boosting is a stage-wise model that minimizes the loss function in hundreds or thousands of weak decision trees, improves predictive performance, and reduces model interpretability.

neuTres is a parameter that provides the threshold for sentence polarity scores to be considered positive or negative and become relevant to this method. Neutral sentences are irrelevant to this method.

orgAcro is a covariate and a unique organization-identifier string for the project's main financier organization: AsDB; IFAD; UNDP; and, WB.

proCode is a covariate and a unique categorical project identifier essential for project record tracking, such as 29586-013 or P050732.

proRate is the overall project rating, the outcome, a textual or numerical value on a formalized six- or four-grade scale ranging from "highly unsuccessful" to "highly successful."

proYear is a covariate that provides the project's end year and the chronological context.

senLeng is a parameter that imposes a minimum sentence length to remove titles, headers, or strings.

simRatio is a parameter that imposes the minimum cosine similarity between the factor and corpus sentence embeddings.

w_i_+ and *w_i_-* are the integer numbers of corpus sentences that are semantically similar to the *i*eth factor and have positive or negative polarity scores, respectively, weighing the factors.

XAI, in this paper, is a sequence of explainable algorithms based on bag-of-words, dictionary, or decision tree ensemble methods without gradient boosting: TF-IDF [50] for embeddings; VADER [51] or TextBlob [52] for sentiment analysis; LR [49] or RF [49], a tree-based ensemble model, for predictions.

ACCESSIBLE DOCUMENTATION

The scripts, hosted on GitHub at https://github.com/nicancies/partnership-prediction-ieeeaccess.git, are available under a GPLv3 license.

ACKNOWLEDGMENT

Nicola Drago conceived, developed, implemented, analyzed, and reviewed the methodology, main script, and simulations, and edited this article in the framework of his Ph.D. research project. Gonçalo Afonso Braz conceived, prepared, implemented, and reviewed all corpus downloading scripts, the main script's Step 4 (sentiment analysis), the re-training of the Stanza model, and project-evaluation-rating matching scripts; he also helped revise the main script's Step 10 (SHAP values). Tullio Vardanega supervised the research project, provided methodological guidance, and edited and reviewed the manuscript. Nicola Drago thanks Dr. Gabriele Orazi of the University of Padova for setting up and assisting with the computing infrastructure and for his patient availability, Prof. Pedro Angel Henriques of the University of Minho for his introduction to and explanations of ontology in computer science, and Prof. José João Dìas de Almeida for his explanations and support in defining

the bulk downloading scripts. The authors acknowledge the use of OpenAI's ChatGPT in laying down this work's main script, following the architecture, specifications, and revisions provided by Nicola Drago, except for Step 4, which was originally prepared by Gonçalo Afonso Braz, following the specifications provided by Nicola Drago.

REFERENCES

[1] United Nations Conference on Trade and Development, *World Investment Report 2022: International Tax Reforms and Sustainable Investment*. New York, NY, USA: United Nations Publications, 2022.

[2] D. R. Boyd. (Aug. 2022). *Report of the Special Rapporteur on the Issue of Human Rights Obligations Relating to the Enjoyment of a Safe, Clean, Healthy and Sustainable Environment, David R. Boyd*. [Online]. Available: https://documents.un.org/doc/undoc/gen/n22/648/97/pdf/n2264897.pdf

[3] Asian Development Bank—Independent Evaluation Department. (2016). *Effectiveness of Asian Development Bank Partnerships*. [Online]. Available: https://www.adb.org/sites/default/files/evaluation-document/155060/files/tes-partnerships.pdf

[4] M. W. McNamara, K. Miller-Stevens, and J. C. Morris, "Exploring the determinants of collaboration failure," *Int. J. Public Admin.*, vol. 43, no. 1, pp. 49–59, Jan. 2020.

[5] N. Drago and T. Vardanega, "Predictive factors for inter-agency partnership success," *Eval. Program Planning*, pp. 1–37, Feb. 2024.

[6] A. Liebenthal, O. N. Feinstein, G. K. Ingram, and R. Picciotto, *Evaluation & Development: The Partnership Dimension*, 1st ed., O. N. Feinstein, Ed., Evanston, IL, USA: Routledge, Feb. 2018. [Online]. Available: https://www.taylorfrancis.com/books/9781351323918

[7] International Fund for Agricultural Development - Independent Evaluation Office. (2018). *Evaluation Synthesis Report on Building Partnerships for Enhanced Development Effectiveness—A Review of Country-level Experiences and Results*. [Online]. Available: https://ioe.ifad.org/en/w/building-partnerships-for-enhanced-development-effectiveness-a-review-of-country-level-experiences-and-results?p_l_back_url=%2Fen%2Fevaluation-synthesis

[8] D. Hollow. (2011). *An Academic Review of the Evaluation of Partnerships in Development*. Accessed: Feb. 6, 2025. [Online]. Available: https://idl-bnc-idrc.dspacedirect.org/server/api/core/bitstreams/95a672b8-a893-49ed-a8b6-ad95d60b2148/content

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (Adaptive Computation and Machine Learning), 1st ed., London, MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[10] *Guidelines for the Evaluation of Public Sector Operations*, Asian Develop. Bank—Independent Eval. Dept., Philippines, Apr. 2016.

[11] *Evaluation Guidelines*, United Nations Develop. Program-Independent Eval. Office, New York, NY, USA, Jun. 2021.

[12] L. G. Morales, Y.-C. Hsu, J. Poole, B. Rae, and I. Rutherford. (Nov. 2014). *A World That Counts: Mobilising the Data Revolution for Sustainable Development*. United Nations, New York, NY, USA. [Online]. Available: https://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf

[13] M. Blanc, T. Esmail, C. Mascarell, and R. Rodriguez, *Predicting Project Outcomes A Simple Methodology for Predictions Based on Project Ratings* (Policy Research Working Paper). Washington, DC, USA: World Bank, Aug. 2016.

[14] P. Amy, J. Craig, and A. Aubra, *Reflecting the Past, Shaping the Future: Making AI Work for International Development*. Washington, DC, USA: USAID, 2018. [Online]. Available: https://2017-2020.usaid.gov/sites/default/files/documents/15396/AIExecutiveSummary-Digital.pdf

[15] O. Dupriez. (2018). *An Empirical Comparison of Machine Learning Classification Algorithms*. [Online]. Available: https://thedocs.worldbank.org/en/doc/666731519844418182-0050022018/original/PRTODpresentationV2.pdf

[16] S. Athey and S. Wager, "Estimating treatment effects with causal forests: An application," 2019, *arXiv:1902.07409*.

[17] D. McKenzie. *How Can Machine Learning and Artificial Intelligence Be Used in Development Interventions and Impact Evaluations?* Accessed: Sep. 15, 2023. [Online]. Available: https://blogs.worldbank.org/en/impactevaluations/how-can-machine-learning-and-artificial-intelligence-be-used-development-interventions-and-impact

[18] International Fund for Agricultural Development. (2020). *Leveraging Artificial Intelligence: Harnessing Big Data to Improve Decision-Making*. [Online]. Available: https://www.ifad.org/en/web/knowledge/-/leveraging-artificial-intelligence-and-big-data-for-ifad-2.0-phase-2

[19] International Fund for Agricultural Development. (2020). *Accelerating Knowledge Generation for Data-driven Decision Making—Leveraging Artificial Intelligence and Big Data for IFAD 2.0*. [Online]. Available: https://www.ifad.org/en/web/knowledge/-/publication/accelerating-knowledge-generation-for-data-driven-decision-making

[20] Asian Development Bank. (2022). *What is ADB's Office of Public-Private Partnership?* [Online]. Available: https://www.adb.org/news/videos/what-adbs-office-public-private-partnership

[21] X. Zhang and C. Cho, "Lessons learned from UNDP support to e-governance: Rapid evidence assessment of evaluations from the last decade," in *Proc. 15th Int. Conf. Theory Pract. Electron. Governance*, New York, NY, USA, Oct. 2022, pp. 552–555, doi: 10.1145/3560107.3560193.

[22] O. A. Garcia, P. Mallari, S. Balakrishnan, and T. Trans. (Mar. 2023). *How the Undp Independent Evaluation O?ce is Using Aws Ai/ml Services To Enhance the Use of Evaluation To Support Progress Toward the Sustainable Development Goals*. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/how-the-undp-independent-evaluation-office-is-using-aws-ai-ml-services-to-enhance-the-use-of-evaluation-to-support-progress-toward-the-sustainable-development-goals/

[23] L. Bravo, A. Hagh, R. Joseph, H. Kambe, Y. Xiang, and J. Vaessen, "Machine learning in evaluative synthesis lessons from private sector evaluation in the World Bank Group," World Bank, IEG Methods Eval. Capacity Develop. Work. Paper Series, Washington, DC, USA, Tech. Rep., Jun. 2023.

[24] *Results and Performance of the World Bank Group 2023*. Washington, DC, USA: World Bank, Dec. 2023. [Online]. Available: https://openknowledge.worldbank.org/handle/10986/40886

[25] V. Kumar, "Supply chain resilience through small and medium-sized enterprises (SMES): Representation, partnership prediction, and standardization using graph databases," Doctor of Philosophy, Dept. Ind. Eng., Pennsylvania State Univ., University Park, PA, USA, Sep. 2023. [Online]. Available: https://etda.libraries.psu.edu/catalog/21006vbk5101

[26] Artifex Software, Inc. (2024). *PyMuPDF*. Python binding for MuPDF. [Online]. Available: https://github.com/pymupdf/PyMuPDF

[27] J. Turk. *Jellyfish: Approximate and Phonetic Matching of Strings*. Accessed: Aug. 20, 2024. [Online]. Available: https://pypi.org/project/jellyfish/

[28] K. Reitz. *Requests 2.32.3*. Accessed: Sep. 1, 2024. [Online]. Available: https://pypi.org/project/requests/

[29] L. Richardson. (Aug. 2024). *Beautiful Soup: We Called Him Tortoise Because He Taught Us*. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/

[30] S. Shafiezadeh, G. M. Duma, G. Mento, A. Danieli, L. Antoniazzi, F. D. P. Cristaldi, P. Bonanni, and A. Testolin, "Methodological issues in evaluating machine learning models for EEG seizure prediction: Good cross-validation accuracy does not guarantee generalization to new patients," *Appl. Sci.*, vol. 13, no. 7, p. 4262, Mar. 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/7/4262

[31] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.

[32] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R* (Springer Texts in Statistics). New York, NY, USA: Springer, 2021.

[33] Microsoft Research—Alice Team. *EconML*. Accessed: Sep. 7, 2024. [Online]. Available: https://econml.azurewebsites.net/

[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[35] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2016, pp. 785–794.

[36] scikit learn. *Neural Network Models (Supervised)*. Accessed: Sep. 14, 2024. [Online]. Available: https://scikit-learn/stable/modules/neuralnetworkssupervised.html

[37] *Mapping the Public Voice for Development—Natural Language Processing of Social Media Text Data—A Special Supplement of Key Indicators for Asia and the Pacific 2022*, Asian Develop. Bank, Philippines, 2022, doi: 10.22617/FLS220347-3.

[38] United Nations Development Program—Independent Evaluation Office. (2023). *(ANCE.501) AIDA—Artificial Intelligence for Development Analytics*. [Online]. Available: https://www.youtube.com/watch?v=WS7p57ayMes
[39] United Nations Development Program. (2024). *Artificial Intelligence for Development Analytics (AIDA)*. [Online]. Available: https://aida.undp.org/landing
[40] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," 2021, *arXiv:2102.10936*.
[41] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.
[42] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. Reading, MA, USA: Addison-Wesley, 1949.
[43] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 2, Aug. 1995, pp. 1137–1143.
[44] United Nations Development Programme. (2024). *Human Development Report 2023–24*. [Online]. Available: http://report.hdr.undp.org
[45] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[46] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2020, pp. 1–8. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf
[47] A. Akbik, D. A. J. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proc. COLING*, Aug. 2018, pp. 1638–1649.
[48] A. Géron, *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA, USA: O'Reilly Media, 2019.
[49] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.
[50] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.
[51] C. Hutto and É. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. Conf. Weblogs Social Media (ICWSM)*, vol. 8, Ann Arbor, MI, USA, May 2014, pp. 216–225.
[52] S. Loria. (2018). *TextBlob: Simplified Text Processing*. Accessed: Aug. 23, 2024. [Online]. Available: https://textblob.readthedocs.io/en/dev/

NICOLA DRAGO received the M.S. degree in mechanical engineering and the Ph.D. degree in computer science for societal challenges from the University of Padova, Italy, in 1994, to harness strategic partnerships and artificial intelligence technologies for more impact on beneficiaries and improved public expenditure efficiency. He has accrued industrial manufacturing experience in corporations and research experience at the Technical University of Denmark on wind energy system components and at the University of Padova on innovation policies. After collaborating as an Expert with the Private Sector Development with multiple UN agencies, multilateral development banks, and bilateral and national development agencies for 20 years in more than 20 countries.

GONÇALO AFONSO BRAZ received the degree in computer science from the University of Minho, Portugal, in 2022, where he is currently pursuing the master's degree, focusing on a thesis about OCR for old, structured documents. In 2024, he was a Research Scholarship with INESC TEC, Porto.

TULLIO VARDANEGA (Member, IEEE) received the M.Sc. degree from the University of Pisa, Italy, in 1986, and the Ph.D. degree from TU Delft, The Netherlands, in 1998. He has been with the University of Padua, Italy, since January 2002. He was PI for a small research and development enterprise, from 1987 to 1991, and then at the European Space Agency, The Netherlands, until December 2001. He specializes in high-integrity real-time systems, edge-to-cloud continuum, software engineering, active learning, and informatics education. He has run several research collaborations in those ambits. He is a member of ACM. He is the Technical Expert for Italy in ISO/IEC JTC1/SC22: WG9 (Ada) and WG23 (Programming Language Vulnerabilities). He was the Chairperson of Ada-Europe for 20 years, from February 2004 to September 2024.