Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

INTERNSHIP REPORT

DATA-DRIVEN DECISION SUPPORT IN PUBLIC PROCUREMENT: A MACHINE LEARNING APPROACH TO CONTRACT AWARD PREDICTION

JAKOB LINNIG

JANUARY - 2026

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

INTERNSHIP REPORT

DATA-DRIVEN DECISION SUPPORT IN PUBLIC PROCUREMENT: A MACHINE LEARNING APPROACH TO CONTRACT AWARD PREDICTION

JAKOB LINNIG

SUPERVISION:

CARLOS J. COSTA

JANUARY - 2026

ACRONYMS

ANN	Artificial Neural Network
API	Application Programming Interface
BI	Business Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
DAX	Data Analysis Expressions
EDA	Exploratory Data Analysis
EFB	Exclusive Feature Bundling
EU	European Union
GDP	Gross Domestic Product
MAE	Mean Absolute Error
ML	Machine Learning
NLP	Natural Language Processing
OCDS	Open Contracting Data Standard
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
PDF	Portable Document Format
Power BI	Microsoft Power BI
RMSE	Root Mean Squared Error
SHAP	Shapley Additive explanations
SME	Small and Medium-sized Enterprise
TED	Tenders Electronic Daily
USD	United States Dollar
URL	Uniform Resource Locator
WB	World Bank

ABSTRACT

Public procurement is economically significant and governance-sensitive, requiring decisions under uncertainty and strong accountability. Although standards define a structured procurement lifecycle and core artifacts (projects, procurement plans, notices and awards), publications are often fragmented, heterogeneous, and only partially linkable across stages. In World Bank procurement, decision-relevant planning information is frequently published as PDF tables rather than as fully structured datasets.

This report designs and evaluates an end-to-end, data-driven decision support solution for World Bank procurement from a supplier (SME) perspective. Using CRISP-DM as a methodological backbone, the artifact integrates multiple procurement sources and implements a document-processing pipeline to extract, clean, and harmonize procurement plan tables from PDFs. The resulting unified analytical layer enables lifecycle-consistent analytics for opportunity screening and competition-oriented exploration.

The modeling task is formulated as supervised regression to estimate contract award values from the integrated dataset. A gradient-boosting approach for categorical tabular data (CatBoost) is applied to handle high-cardinality categorical variables and heterogeneous feature spaces typical of procurement data. To support stakeholder trust and validation, model outputs are complemented with explainability using SHAP, enabling interpretation of feature contributions at both global and case level.

Evaluation emphasizes predictive performance relative to transparent baselines, robustness across procurement segments, and the practical usefulness of explanations for screening and prioritization decisions. The solution is deployed through a Power BI dashboard that consolidates lifecycle information, predicted award values, and procurement plan views in one interface. A scheduled refresh pipeline automates extraction, preparation, and scoring, requiring only a dataset refresh by the user and improving timeliness while reducing manual workload. The implementation serves as a template for comparable contexts.

Overall, the report demonstrates how document-based procurement disclosures can be transformed into an integrated decision-support layer and how explainable machine learning can enhance evidence-informed tender screening and analysis in an applied World Bank procurement setting.

ABSTRACT PORTUGUESE

A contratação pública é economicamente significativa e sensível em termos de governação, exigindo decisões sob incerteza e forte responsabilização. Embora as normas definam um ciclo de vida estruturado e artefactos centrais (projetos, planos de contratação, avisos e adjudicações), as publicações são frequentemente fragmentadas, heterogéneas e pouco ligáveis entre fases. No Banco Mundial, a informação de planeamento é muitas vezes publicada em tabelas PDF, em vez de conjuntos de dados estruturados.

Este relatório concebe e avalia uma solução de apoio à decisão, orientada por dados, para a contratação pública do Banco Mundial, na perspetiva de um fornecedor (PME). Seguindo a metodologia CRISP-DM, o artefacto integra várias fontes de contratação e implementa um pipeline documental para extrair, limpar e harmonizar as tabelas dos planos de contratação a partir de PDFs. A camada analítica resultante permite análises consistentes ao longo do ciclo de vida, orientadas para o rastreio de oportunidades e a análise da concorrência. A modelação é formulada como regressão supervisionada, para estimar valores de adjudicação a partir do conjunto de dados integrado. Aplica-se uma abordagem de gradient boosting (CatBoost) para lidar com variáveis categóricas de elevada cardinalidade, típicas destes dados. Os resultados são complementados com explicabilidade via SHAP, permitindo interpretar os contributos das variáveis a nível global e individual.

A avaliação incide sobre o desempenho preditivo face a baselines transparentes, a robustez entre segmentos e a utilidade prática das explicações para decisões de rastreio e priorização. A solução é disponibilizada num dashboard Power BI que consolida informação do ciclo de vida, valores previstos e planos de contratação numa só interface. Um pipeline de atualização agendada automatiza extração, preparação e scoring, exigindo apenas uma atualização de dados pelo utilizador melhorando a tempestividade e reduzindo o trabalho manual. A implementação serve de modelo para contextos semelhantes.

Em suma, o relatório mostra como divulgações documentais de contratação pública podem tornar-se numa camada integrada de apoio à decisão, e como a aprendizagem automática explicável pode reforçar o rastreio e a análise de concursos baseados em evidência, num contexto real do Banco Mundial.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

Public procurement is a high-impact domain in which decisions combine substantial economic relevance with strict accountability requirements. In OECD member states, procurement is often reported at around 13% of GDP, implying that even small improvements in how procurement information is processed and acted upon can have meaningful fiscal and societal effects (Obwegeser and Müller, 2018). At the same time, procurement outcomes are closely linked to transparency and integrity concerns, and procurement data has been used to derive indicators for monitoring and oversight (Fazekas and Kocsis, 2020). This makes procurement a natural setting for data-driven decision support, provided that heterogeneous disclosures can be transformed into reliable analytical inputs.

In practice, this transformation is difficult. Although standards such as the Open Contracting Data Standard (OCDS) define a procurement lifecycle and core artifacts (tender, award, contract), disclosures are often fragmented, heterogeneous, and only partially linkable across stages (*Open Contracting Data Standard*, n.d., Fazekas et al., 2024). Large-scale procurement datasets show systematic missingness and incomplete lifecycle linkage for important reference variables, limiting end-to-end analysis from planning to award outcomes (Csáki and Prier, 2018). These problems are amplified when decision-relevant information is published in document form rather than as structured data. In the World Bank procurement context, planning information is frequently disclosed as PDF tables, creating additional extraction, harmonization, and data-quality constraints before analytics and modeling can be operationalized.

Recent work suggests that machine learning can support procurement decision making through monitoring and decision-support applications and that ML models can predict procurement outcomes such as prices or values under suitable data conditions (Nai et al., 2023, Kim and Jung, 2018, García Rodríguez et al., 2019, 2021). Yet deployable procurement decision support must go beyond predictive accuracy. Procurement data is strongly categorical and heterogeneous, requiring robust tabular modeling approaches, and the procurement context demands interpretable outputs to support trust, validation, and accountable use (Lundberg and Lee, 2017, Prokhorenkova et al., 2019).

Against this background, this work designs and evaluates an end-to-end, data-driven decision support solution for World Bank procurement from a supplier (SME) perspective. The guiding research question is: *Can machine learning be effective in decision making in public procurement and lead to more success?* Effectiveness is evaluated through predictive performance beyond simple baselines, robustness under heterogeneous and noisy disclosure conditions, and interpretable explanations suitable for practical stake-

holder use (Lundberg and Lee, 2017, Prokhorenkova et al., 2019). Concretely, the work targets four objectives that are carried through the remainder of the report. Consolidation of heterogeneous World Bank procurement artifacts (projects, procurement plans, notices, and awards) into a unified analytical layer, delivering early-stage award-value estimates for upcoming opportunities via supervised regression on the integrated dataset, complementing predictions with stakeholder-oriented interpretability to support transparency and reviewability and reducing manual effort by using the pipeline outputs in a BI setting with regular refresh for ongoing use.

Methodologically, the work follows a Design Science Research approach and uses CRISP-DM to document the progression from business understanding to deployment and to support iterative refinement driven by data constraints (Costa and Aparicio, 2020, Chapman, 2000, Wirth and Hipp, 2000, Schröer et al., 2021). Overall, the report aims to bridge procurement data heterogeneity and document-based disclosure realities with stakeholder-facing analytics and explainable machine-learning-based estimation in an applied World Bank procurement setting.

2 LITERATURE REVIEW

This chapter reviews research and standards relevant to procurement decision support, procurement data quality, and machine learning for award value prediction. The purpose is to establish the theoretical and empirical basis for the methodological choices adopted in this project, and to position the work against persistent practical constraints in real-world procurement data ecosystems.

To connect the literature to the study objective, the review is organized around the specifications an end-to-end decision support solution must provide: decision support value in procurement, the procurement lifecycle and its key artifacts as information inputs, data ecosystem and quality constraints that shape feasible analytics, and finally, predictive modeling and explainability methods. This structure enables an explicit synthesis of what prior work already solves, what remains difficult in heterogeneous procurement settings, and why an integrated pipeline is a defensible response to that gap.

2.1 Public Procurement and Data-Driven Decision Support

Public procurement has evolved from an administrative function into a strategic policy governance domain in which data-driven analysis and decision support can improve efficiency, transparency, and accountability for all parties involved. This shift is motivated by the growing amount of procurement activity in the public sector. Obwegeser and Müller (2018) highlight that OECD member states spend on average 13% of GDP on public procurement, accounting for a substantial part of total government expenditures. At this scale, even a slight improvement in efficiency, competition, or risk control can have a tremendous fiscal and societal impact (Obwegeser and Müller, 2018).

A key motivation for data-driven decision support from an institutional perspective is the role of procurement as a governance and integrity risk area. Fazekas and Kocsis (2020) show how procurement data can be used to build objective proxy measures for high-level corruption, operationalizing mechanisms such as unjustified restriction of access to public contracts. Using large-scale contract-level data across multiple European countries, they argue that procurement data can provide actionable indicators that complement perception-based governance measures and directly support monitoring and policy interventions (Fazekas and Kocsis, 2020). This work underscores two important points for procurement analytics: first, that procurement data can be transformed into meaningful risk and performance indicators, and second, that decision support is not limited to forecasting or optimization.

Recent applied research further shows that machine learning can support deci-

sion makers in procurement settings with concrete and operationally relevant outcomes. Nai et al. (2023) describe how the growing availability of e-procurement platforms and open procurement data enables the joint use of multiple information sources. They train machine learning models to identify procurement procedures that are likely to lead to disputes and complaints, and they develop a recommender system to find similar procurement cases to assist stakeholders in navigating those procurement environments (Nai et al., 2023). Their results indicate that data-driven systems can serve as practical decision support, for example by flagging tenders with higher risk or showing comparable cases that support consistent handling of procedures. They also emphasize a recurring challenge in public-sector machine learning: decision support systems are often perceived as "black boxes," which increases the demand for trustworthy and explainable models in procurement contexts where decisions are subject to scrutiny (Nai et al., 2023). This strengthens the argument that model transparency and interpretability are not optional, but rather core design requirements for procurement decision support.

The relevance and value of data-driven procurement decision support becomes even clearer with the emergence of standardized datasets that are supposed to enable analysis at scale. Fazekas et al. (2024) introduce a global contract-level public procurement dataset that addresses the lack of globally comparable data on tenders and awards. Their work highlights both opportunities and challenges: procurement spending is very high at a global level, yet analytics across countries and institutions are often limited by data fragmentation and inconsistent reporting. Harmonizing procurement data to a common standard enables cross-sectional analysis of procurement markets, spending patterns, and integrity risks (Fazekas et al., 2024).

All of the above provide evidence that the public procurement sector is attractive for data-driven decision support due to the scale and economic value, its institutional emphasis on traceability and accountability, and its potential for governance-oriented analytics. These characteristics motivate the main focus of this work by applying machine learning and analytics to support decision makers from a company's point of view that is competing in public procurement.

2.2 Procurement Lifecycle and key artifacts

According to *Open Contracting Data Standard* (n.d.), a procurement process can be simplified and described by a sequence of stages. Each of these stages has key artifacts and documents that are vital for the entire process (*Open Contracting Data Standard*, n.d.).

FIGURE 1: OCDS process sequence. Source: (*Open Contracting Data Standard*, n.d.).

OCDS, or Open Contracting Data Standard, is an initiative to provide a standard in public procurement for which key documents and data should be published in each stage of the procurement process (*Open Contracting Data Standard*, n.d.). In the context of this work, which focuses only on World Bank procurement, these stages are also observable. According to the World Bank (2025) Procurement Regulations, in the planning stage the borrower is obliged to prepare a procurement plan prior to negotiations with the Bank. If the Bank approves it, the project moves to the next stage.

Procurement plans contain the planned activities for the entire project with key attributes such as the selection methods, evaluation approach, cost estimates, etc. (World Bank, 2025). They are the central point of the procurement process. When the process moves along the pipeline, the borrower has to provide more detailed documents about each planned activity. This procurement structure led to the specific data choice for this work. The focused data sources are the procurement plans, notices, contract awards, and project metadata. From a decision-support perspective, these artifacts are complementary information inputs. Consolidating them into a unified analytical layer can reduce search effort while enabling consistent comparisons across opportunities.

2.3 Data Ecosystems and Data Quality Challenges

Public procurement data, even with OCDS, still suffers in terms of data quality. The data is produced in a data ecosystem with multiple actors like contracting authorities, central portals, or secondary users that all produce data. Fazekas et al. (2024) argue that heterogeneity does not only arise from institutional differences and implementation, but also from technical publication choices and the fragmentation of data across multiple systems and processes. They therefore conclude that procurement datasets require an extensive preprocessing phase before performing any analysis (Fazekas et al., 2024). Fazekas et al. (2024) also state that a key driver for heterogeneity in procurement data is the inconsistency of publishing formats across and even within institutions, which makes creating a global contract-level dataset difficult.

Heterogeneity and fragmentation are not the only challenges when dealing with

procurement data. An empirical study performed by Csáki and Prier (2018) based on TED (Tenders Electronic Daily), a European Union procurement platform, indicates high complexity in files due to embedded levels of procurement information and highlights a range of issues with non-normalized fields, missingness of data, and lack of lifecycle linkage. They find that fields, and especially key reference variables, are often missing, with a 66% missingness rate across a consolidated period, which makes complete tender-to-award linkage very difficult (Csáki and Prier, 2018).

Taken together, the literature suggests that data quality challenges are systematic properties of the procurement system, caused by heterogeneous publication formats, fragmented data across one institution or multiple institutions, and missing data, especially for reference variables (Csáki and Prier, 2018, Fazekas et al., 2024). These challenges imply that an extensive data preparation stage is not optional but rather a key foundation step for any analytical approach in public procurement. This aligns with the extensive preprocessing steps taken in this work prior to applying machine learning and analytics. Therefore, reducing manual effort and enabling reliable analytics in procurement is not a matter of selecting sophisticated models, it is a matter of systematically addressing heterogeneity, missingness, and weak linkage across lifecycle stages (Csáki and Prier, 2018, Fazekas et al., 2024).

2.4 Machine Learning in Public Procurement

The growing access to open data from public procurement combined with the increasing adoption of e-procurement systems are enabling the use of machine learning for various different use cases like analytics, decision support, or governance. Nai et al. (2023) emphasize that the increased use of e-procurement systems in the public sector enables joint access to different information sources, leading to the possibility of machine-learning-based approaches in public procurement rather than purely academic experiments.

In the research, three main areas of machine learning application emerged: monitoring, market analysis, and predictive decision support. Monitoring refers mainly to the use of machine learning to monitor governance and integrity risks. Fazekas and Kocsis (2020) propose objective proxy measures for corruption risks which are derived from procurement data. They define corruption in that case as an unjust denial of access to public contracts in favor of a different bidding party. A more predictive approach in the area of monitoring is presented by Nai et al. (2023) who combined Italian procurement data with judicial records with the aim of identifying procurement features that are associated with anomalies. This approach shows that machine learning can be used in a predictive setting

for public procurement data.

Beyond monitoring, ML is also used to reduce information asymmetries in the procurement sector between supplier and contractor and increase competition. Nai et al. (2023) not only predicted tenders that could lead to anomalies but also developed a recommender system that finds similar procurements for a given one. Another recommender approach is shown by García Rodríguez et al. (2019) in their work, which aims to allow the contracting authority to discover companies that are suitable for a chosen tender. Methodologically, they implement a random forest classifier for the recommendation system. Their hypothesis is that tenders with more competition will have lower prices and that competition will make corruption more difficult. Their results show that the recommender system acts as an enabling mechanism for more market access, especially for SMEs, which facilitates more competition and less corruption (García Rodríguez et al., 2019). The literature leads to the conclusion that recommender systems are not convenience tools, but can be mechanisms for increased competition and enhanced transparency objectives, as well as showing that ML can have strong application scenarios in the procurement market.

The third major application area deals with the prediction of quantitative outcomes such as award prices and bid levels. These predictions can be valuable for economic reasons and deliver predictions that are insightful and outperform baseline approaches. In an analysis of Spanish procurement announcements, García Rodríguez et al. (2019) show that the tender price is not a reliable proxy for the final award price by showing a clear difference between percentage errors and different mean and median behaviors. They argue that an accurate award price estimator would be valuable and would reduce economic risks and that machine learning has been chosen due to the complexity of the prediction problem. They also show the dependency of predictive modeling in procurement on data preparation (García Rodríguez et al., 2019).

A later study by García Rodríguez et al. (2021) evaluates multiple different ML approaches for award price prediction. They argue that ML price predictions are a relevant research topic to help address the problem of inefficient use of public resources in procurement. In the study, they compare the following ML models: linear regression, random forest, isotonic regression, and neural network models. They provide implementation details and evaluate the models with multiple error metrics. They use a random forest as the baseline and show that isotonic regression and ANNs can outperform the baseline for award price predictions (García Rodríguez et al., 2021).

Kim and Jung (2018) provide complementary evidence for price prediction in procurement settings by discussing how empirical analysis often suffers from multicollinear-

ity and high dimensionality. They apply random-forest variable selection with regularized regression. They report evidence that the approach provides better results relative to an ordinary least squares benchmark (Kim and Jung, 2018). Collectively, these studies show that procurement ML is not limited to classification (e.g., monitoring risk), but also covers robust regression and feature-selection strategies for modeling price and value outcomes. This motivates the chosen approach for this work, by applying a CatBoost regression to predict contract award values in the public procurement setting.

A lot of procurement data is tabular and includes high-cardinality categorical variables (e.g., country, regions, procurement method, procurement sector). Gradient-boosted decision trees are, as the literature before showed, widely used in such settings. One of those gradient-boosted decision tree algorithms is CatBoost. CatBoost, as the name suggests, focuses on categorical features, the mitigation of target leakage, and handling of missing values by proposing ordered boosting and a dedicated approach to categorical feature processing (Prokhorenkova et al., 2019).

For public-sector deployment, we saw that results are important, but explainability of how the results were achieved by the model is equally important. Lundberg and Lee (2017) argue that understanding why a model makes a prediction can be equally valuable as the prediction itself, and introduce SHAP as a framework in which SHAP assigns a feature importance value to each individual feature for a given prediction value. Nai et al. (2023) operationalize this principle by adopting SHAP as an explanation method in their procurement setting.

In the area of public procurement, the main features for prediction are of a categorical nature, which led to the use of CatBoost as the main algorithm of this work. Additionally, the other properties of CatBoost, like the mitigation of target leakage and handling of missing values, further solidified the design choice.

Finally, evaluation choices are non-trivial and should be justified: Botchkarev (2019) emphasizes that performance metrics are vital components of evaluation frameworks and proposes a typology to guide metric selection in regression and forecasting. Hodson (2022) likewise notes a continuous confusion about the use of RMSE versus MAE and highlights that neither metric is universally superior to the other, with each corresponding to different implicit error assumptions. In procurement contexts, where the costs of over- versus underestimation may differ, this metric reasoning becomes particularly important. Therefore, both Botchkarev (2019) and Hodson (2022) motivate selecting evaluation metrics carefully and in line with their work.

In summary, the reviewed literature implies three requirements for the stated objectives of this work. Decision support must be operationally useful and trustworthy in

high-scrutiny public-sector contexts (Nai et al., 2023). Analytics must be grounded in the procurement lifecycle and its key artifacts (*Open Contracting Data Standard*, n.d., World Bank, 2025). And modeling performance depends on robust data preparation that mitigates heterogeneity, missingness, and weak linkage (Csáki and Prier, 2018, Fazekas et al., 2024). The next section shows how these requirements expose a gap between prior research assumptions and end-to-end deployment realities.

2.5 Research Gap and Positioning of the Present Work

In practice, procurement data is mostly not available in clean, machine-readable form. Open procurement data can be inconsistent and incomplete, making substantial data quality work inevitable before any analytics is possible (Csáki and Prier, 2018). This is amplified by the fact that data is stored in documents. Key information frequently resides in PDFs and tables, where extraction is error-prone and tool performance varies strongly by layout and document type (Adhikari and Agarwal, 2025).

Existing ML work shows that procurement outcomes such as award price or value can be predicted, but these studies typically assume clean data and are often constrained to single-country contexts and do not show a complete pipeline from data collection to decision support, limiting transferability to heterogeneous data ecosystems (García Rodríguez et al., 2021). Moreover, explainability is rarely integrated into stakeholder-facing delivery, despite the importance of transparent model behavior for trust; SHAP provides a principled way to explain individual predictions for complex models (Lundberg and Lee, 2017).

This work addresses the gap by showing a complete pipeline from World Bank style procurement artifacts to actionable decision support: extracting and harmonizing data from multiple sources, training a robust model for award value estimation (CatBoost), explaining outputs with SHAP, and operationalizing results in a BI environment following an end-to-end project framework.

3 METHODOLOGY

This work adopts a Design Science Research (DSR) perspective. In DSR, the central contribution is an artifact that addresses a practical decision support problem through a rigorously documented design and evaluation process (vom Brocke et al., 2020). To provide a clear, reproducible, and widely recognized structure for this end-to-end machine learning approach, the methodological foundation follows the CRISP-DM process model.

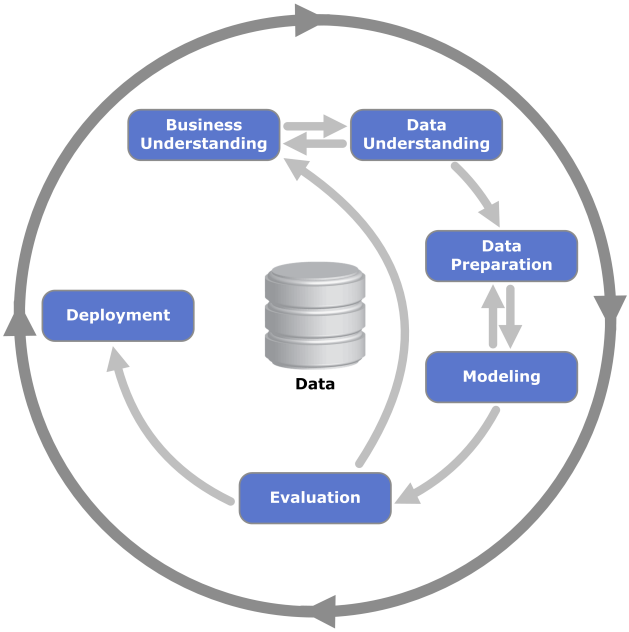

FIGURE 2: CRISP-DM process model. Source: (Jensen, 2012).

CRISP-DM was proposed as a comprehensive, cross-industry and technology-agnostic standard for data mining projects, aiming to make analytical work comparable and executable across domains and tools (Wirth and Hipp, 2000). In its reference model, CRISP-DM organizes the project lifecycle into six phases business understanding, data understanding, data preparation, modeling, evaluation, and deployment while explicitly emphasizing that effective projects are iterative rather than strictly linear and often require movement back and forth between phases as intermediate results reveal new constraints or opportunities (Chapman, 2000). This iterative property is particularly suitable in procurement analytics, where heterogeneous sources, missingness, and weak linkage across lifecycle stages frequently necessitate repeated cycles of data understanding and preparation before robust modeling is feasible.

The continued relevance of CRISP-DM in applied research is supported by methodological literature: Schröer et al. (2021) highlight that CRISP-DM remains a standard approach in practice, while also noting that deployment is often underrepre-

sented in academic reporting. Complementing this point, Costa and Aparicio (2020) process-oriented methodology review (POST-DS) positions CRISP-DM as one of the most used approaches, while stressing that deployment and the presentation of results are critical for translating models into usable decision support (Costa and Aparicio, 2020). Taken together, these arguments justify CRISP-DM as an appropriate backbone for an objective-driven, end-to-end decision support solution. It offers a well-established phase structure, explicitly accommodates iteration driven by data constraints, and keeps deployment and stakeholder-facing delivery within the methodological scope rather than treating them as an afterthought.

Accordingly, the remainder of this chapter documents the design and implementation of the solution along the CRISP-DM phases (Figure 2), describing how business objectives are translated into analytical tasks, how data is collected and transformed into a unified schema, how the model is trained and evaluated, and how outputs are utilized to support real procurement decision processes.

3.1 Business Understanding

This work was developed from a problem paper from stakeholders. The domain of public procurement suffers from heterogeneous data, and so do the stakeholders. In this case, this work was developed for an SME in the domain of software development and consulting, specifically with the help of the international sales team. The main focus of this team is public procurement and specifically multilateral procurement via the World Bank.

The main driver for this work is the fragmentation of data across multiple sources even within single entities like the World Bank. The main artifacts that are relevant for the team are project information, procurement plans, and contract awards. These are scattered across multiple sources, which makes navigating from source to source to obtain the information you are looking for tedious. This information is needed to make informed decisions about tender participation, reducing information asymmetry and increasing the chance of winning tenders.

Based on this problem description, this work turned the problem into a well-defined decision support task with objectives, constraints, and success criteria from the perspective of a competing organization. The main goal is improving decision making to increase the possibility of winning procurements, by unifying heterogeneous and fragmented data into a single space and applying predictive ML regression to predict contract award values for better cost estimation and competition insights. This project is also intended to drastically reduce the manual effort of information search and improve decision

making capabilities, guiding the stakeholder and not replacing them.

Based on the problem specification above, the work is mainly used as a predictive modeling task and secondarily as a data collection and visualization task. The model's purpose is to predict the future value of a possible contract award for a new sub-project at the time of publication of said sub-project, by using a set of features and past contract award notices. The output of the model is a prediction and should never be treated as a fixed value, but more like an estimate.

By applying this ML approach, the work is trying to answer the following research question:

1. *Can machine learning be effective in decision making in public procurement and lead to more success?*

To measure the results of the ML approach, the work utilizes multiple metrics and methods for predictive accuracy, explanation capabilities of the model, comparison against a simple baseline, and robustness. Complementary to evaluation, there are also constraints and risks related to the work. The work is constrained by heterogeneous and document-based publication from the entity that may require special extraction and normalization steps before modeling, incomplete and incorrect data, and access level restrictions to data, since not all data is openly available. The main risks for the ML model are data leakage and selection bias, which means that the observed outcome is not fully representative of the whole data. These risks are addressed by leakage-aware feature selection, careful split strategies aligned with operational use, and error analysis.

3.2 Data Understanding

From the information obtained from the business understanding and the problem description of stakeholders, the important data for this work are structured datasets from the World Bank as well as project metadata, procurement plans, notices, and contract awards. The data is dependent on each other and illustrates the information and data flow through the World Bank procurement process.

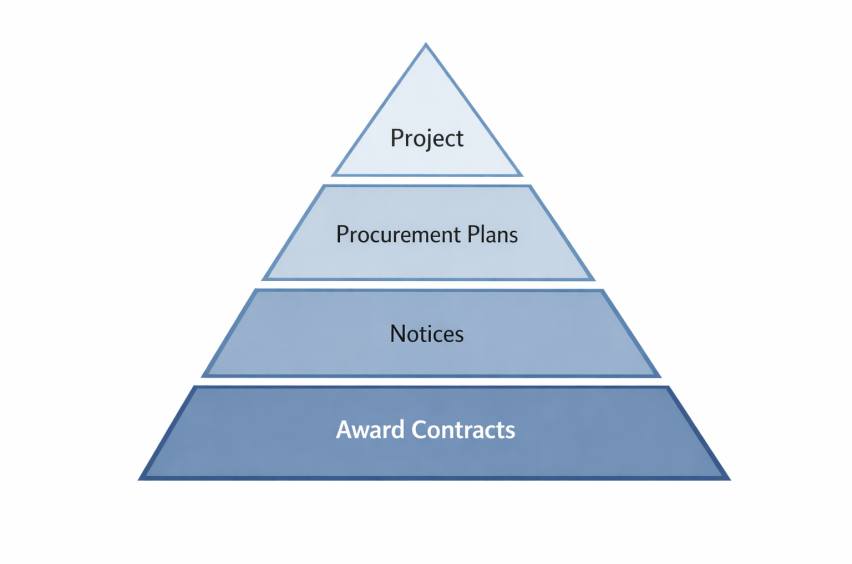

FIGURE 3: Procurement artifact hierarchy. Source: (OpenAI, 2026)

The project data contains metadata for the entire project scope with basic information about the entire project. The core part of the World Bank procurement process are the procurement plans. These contain all the activities that are planned for the project and are available as text files and PDF files. All those activities are shown in tables, so this information has to be extracted from the PDF files, since the text files are not aligned and do not have a proper parsing format. Therefore this work downloads the PDF files and extracts the table information from these. The notice data are activities from the procurement plan that are now officially published and the award contract data are notices that have moved forward to a supplier and borrower agreement.

The notice and award contract data is available via the World Bank Data Catalog API and is fetched completely without any type of date or geographic filtering. This data is openly available via the Data Catalog API with specific dataset and resource IDs. The project data is collected via the World Bank Search API and prefiltered by removing columns where more than 50% of values were missing. This step was implemented after exploring the raw data and evaluating the applied filter. For the procurement plans, the document metadata was also extracted via the World Bank API. The data was filtered to only extract procurement plan documents and after extraction, with the help of the URL column, the most recent procurement plan PDF was downloaded.

This extraction process delivers Parquet files to store the raw data frames for project metadata, notices, and contract awards, and a folder structure with all the procurement plans. Each dataset has the `ProjectID` as a column, making linking to the project easy, but relating the entire way from project to award is not possible, since linking

variables are missing. This granularity mismatch was a challenge especially for feature selection and feature engineering and was overcome by aggregating on different features that are possible to link, like `ProjectID` or `procurement method`.

After extraction, the dataset contains $22{,}729$ projects, over one million rows of procurement plans, $227{,}781$ notices, and $208{,}190$ contract awards. The first exploratory data analysis showed that missingness is a major problem, along with the data being heavily skewed, which was to be expected with procurement data due to the nature of the data. On top of that, there were multiple inconsistencies in naming and formats across the data. The EDA indicated that a thorough and iterative approach between data understanding and data preparation was necessary and a key step in the work's process.

3.3 Data Preparation

Based on that the data preparation pipeline of this project was carefully developed and iteratively improved. In this work the preparation is not limited to standard preprocessing steps like type casting or missing-value handling, but additionally performing document based extraction from procurement plan PDFs, schema harmonization across artifacts and sources as well as the integration into a unified analytical layer aligned with the decision-support objective. Given the strong heterogeneity observed during the exploratory analysis, the preparation workflow was implemented as a reproducible, multi-stage pipeline with intermediate persistence to enable iterative refinement and traceability.

A central challenge of the World Bank procurement plan data is that key information is published as PDF tables with varying layouts and page structures. To make these artifacts useful for creating any analytics, procurement plan PDFs were first filtered to reduce non-informative and computationally expensive cases. This included heuristics based on file size and document recency, and the selection of the most recent plan per project where filename conventions allowed a ranking to be inferred. Candidate pages containing large tabular structures were then detected via a lightweight geometric heuristic (counting lines and rectangles per page). Table extraction was performed using a lattice-based parser and executed in parallel across documents for scalability. The filtering and usage of a lattice-based parser are due to the nature of pdf parsing being computational expensive and the structure of the table inside the pdf, which are tables with lines that a lattice-based extractor scans for and then extracts the content based on those horizontal and vertical lines. The Extracted tables were stored as individual Parquet files and organized in a project-centric folder structure to preserve project structure and enable subsequent per-project cleaning and inspection.

Raw extractions from PDFs frequently contain tables spreading across multiple

pages, repeated header blocks and misalignment of data inside the tables. Therefore, a dedicated cleaning pipeline was applied to the extracted Parquet tables. First, tables were filtered by minimum column count to exclude spurious detections. Second, tables spanning multiple pages were stitched into one based on the section label rows, reflecting the common procurement-plan structure where a section indicates the start of a new table block. Third, multi-row headers were flattened into a single canonical header row by combining main headers with subheaders, including forward-filling header cells to correct for blank header positions. Fourth, systematic text normalization addressed PDF-specific noise such as newline splits, non-breaking spaces, and inconsistent whitespace. Finally, planned/actual schedule columns were parsed into datetime types where feasible, and monetary columns were normalized by removing separators, coercing to numeric values, and treating structural zeros as missing values. Cleaned tables were then combined and stored as Parquet to preserve a clear boundary between extraction and normalization. The result of this step is a single, schema-consistent procurement-plan dataset that can be joined and aggregated for feature engineering.

In parallel to the document-based pipeline, the structured datasets (projects, notices, and contract awards) were standardized to support consistent joins. This included the removal of high-missingness or analysis-irrelevant attributes, the elimination of patterned meaningless columns, and systematic renaming and typing of identifiers to enforce a single `ProjectID` key representation across sources. Numeric measures data types were changed to numeric with invalid values becoming missing instead of doing any sort of imputation, and unit/scale adjustments were applied where required to ensure comparability of magnitude across financial fields. The cleaned and standardized datasets were saved as Parquet files to serve as the stable input for feature engineering, modeling, and BI deployment.

The output of the data preparation phase is a reproducible set of curated Parquet datasets (projects, notices, awards, and unified procurement plans) that share consistent identifiers and normalized variable representations. However, while `ProjectID` enables integration at the project level, end-to-end linkage from project $\rightarrow$ notice $\rightarrow$ award is not fully supported for all records due to missing linkage variables that would allow row-wise linkage. Consequently, subsequent feature engineering operationalizes predictors using aggregations at joinable levels (project-level or method-level summaries), thereby preserving coverage while mitigating linkage sparsity. This preparation design supports the later CRISP-DM phases by ensuring that modeling and evaluation operate on traceable, schema-consistent inputs rather than on heterogeneous raw artifacts.

3.4 Modeling

The modeling phase translates the prepared analytical layer into a predictive model that can be evaluated and later deployed as part of the decision support artifact. In this work, modeling is formulated as a supervised regression task with the objective of estimating the monetary value of a future contract award based on information available at decision time and patterns learned from historical contract award outcomes.

The dependent variable in this work is the award-level contract amount in USD (`supplier_contract_amount_usd`). The dataset used for the modeling of the ML algorithm is created by enriching the award data with possibly informative variables from project metadata and procurement-plan-derived aggregates. In particular, estimated amount aggregations extracted from procurement plans from multiple granularities (e.g., by `ProjectID`, `procurement_category`, and `procurement_method`) are calculated and joined to award and notice observations to yield a consistent feature space for training and inference.

The feature set combines categorical descriptors (e.g., `country_name`, `region_code`, `procurement_category`, `procurement_method`) with numeric aggregates that summarize planned amounts from procurement plans. As a result of the data being skewed and to stabilize relationships in monetary variables, log-transformed variants of selected aggregate features are computed using a guarded `log1p` transformation (stored with a `log1p_` prefix). Missing categorical values are mapped to a dedicated `Unknown` label, and numeric columns are coerced to numeric types to ensure stable ingestion by the learning algorithm.

Procurement records are nested within projects and share common context (e.g., budgets, locations, and planning conventions). To prevent data leakage due to the aggregation on project level, the split strategy is defined at `ProjectID` level. Concretely, unique projects are partitioned into training and test sets, and all award rows belonging to a given project are assigned to the test or training data. To be sure a check was performed to ensure that no `ProjectID` is shared between partitions.

Model training uses gradient-boosted decision trees implemented via CatBoost. CatBoost is suitable for heterogeneous procurement data because it natively supports categorical predictors alongside numerical features without requiring extensive one-hot encoding and handles missing data well. To address the heavy-tailed distribution of contract values, the target variable is modeled on a log scale using $\log(1 + y)$. The predictions are transformed back using the inverse function $\exp(\hat{y}) - 1$.

The modeling phase produces deployable outputs beyond the fitted prediction.

This includes model files, the final feature list (including categorical column indices), and prediction outputs for the test set to support subsequent error analysis and interpretability. These outputs form the base for the evaluation and deployment phases and enable applying the trained estimator to notice-level feature tables to generate award value estimates.

3.5 Evaluation

The evaluation phase of this work focuses on predictive performance for contract award value estimation, robustness across heterogeneous procurement segments, and transparency through explanations that support stakeholder trust.

Contract values are heavy-tailed and exhibit strong right-skew, which makes point estimates sensitive to rare, extremely large contracts. Having this in mind, the performance of the model is evaluated in both the original monetary scale (USD) and a logarithmic scale. Let y denote the observed contract amount and $\hat{y}$ the predicted amount. The primary error metric is the mean absolute error (MAE),

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \,, \tag{1}$$

supplemented by the root mean squared error (RMSE) and the coefficient of determination (R^2). In addition, the same metrics are computed in $\log(1 + y)$ space to reduce the dominance of extreme values and to better reflect proportional deviations for very large awards. This is consistent with the overall preprocessing strategy, which introduces log-transformed features for aggregated monetary estimates (e.g., `log1p_proj_est_mean`, `log1p_meth_est_median`).

To interpret the impact of the learned model, results are compared against simple, transparent baselines. A global baseline that predicts a single central tendency of historical awards for all observations and a category-conditional baseline that predicts the historical central tendency within `procurement_category`.

These baselines serve as lower bounds for expected performance and help determine whether the model adds some insights compared to when we would just consider the median of all contracts as a prediction.

In addition to performance metrics, visualizations are used to check for systematic error patterns and to understand where the model is reliable and explainable.

Residual plots ($\log(1+y) - \log(1+\hat{y})$ vs. $\log(1+\hat{y})$) are inspected for heteroscedasticity and structural deviations. Figure 4 illustrate these diagnostics.

Because decision usefulness differs by contract size (e.g., high-value awards have

disproportionate strategic relevance), MAE is additionally analyzed across deciles of the true contract value. This decile-based evaluation highlights whether the model performs uniformly or whether errors concentrate in the highest-value tail. Figure 5 reports MAE by decile.

Given the heterogeneity of World Bank procurement data, robustness is assessed across key categorical dimensions, particularly `procurement_category`. This dimension is also central to the feature construction step, where procurement-plan estimates are aggregated by category. Figure 6 summarize MAE by the most frequent categories.

Since the project is supposed to help stakeholders in decision making the model transparency and understandability is a key factor and evaluation criterion. Prediction explanations are generated using SHAP values, which attribute the prediction to individual features for both global and local interpretability.

A SHAP summary bar plot ranks features by mean absolute contribution, providing a high-level view of which variables drive predictions across the test set. This is used to validate whether the model relies on plausible signals (e.g., procurement method and historical plan-derived estimates) rather than spurious artifacts. Figure 7 shows the resulting global ranking.

Finally, local explanations are used to analyze individual predictions, with a focus on cases with large absolute error. A SHAP waterfall plot decomposes a single prediction into additive feature contributions relative to the expected value, showing which attributes are driving the estimates in what direction and allow for analysis of those values. Figure 10 provides a worst-error example, illustrating how the model combines categorical context (e.g., `procurement_method`, `country_name`) with aggregated monetary estimates.

The evaluation phase provides evidence on whether the model is suitable as a decision support component under the constraints of heterogeneous, partially missing procurement data. Observed error concentrations in specific deciles or procurement segments directly inform iterative refinement steps in data preparation. In this sense, evaluation acts as a controlled feedback loop that links model performance back to the data and feature assumptions established in earlier CRISP-DM phases.

3.6 Deployment

Once the iterative process of evaluating and modeling is done, the solution is deployed as a two-layer artifact. An automated, scheduled data and scoring pipeline that produces refreshed analytical tables and model predictions, and a stakeholder-facing

Power BI dashboard that unifies project information, procurement plans, notices, contract awards, and predicted award values in a single decision support space. The primary design goal is to minimize manual effort and technical friction for the end user. Once the pipeline has executed, the user only needs to refresh the Power BI dataset to access the latest integrated information and predictions.

Model inference is implemented as a batch-scoring step that applies the trained and saved model to the prepared notice level feature table. Concretely, the pipeline loads the model file and a metadata file that specifies the exact feature columns and categorical columns used during training. It then reads the prepared notice dataset (`notices_ML.parquet`), selects the required feature subset, performs prediction, applies the inverse transformation from log-space via `expm1`, and appends the resulting prediction as an additional column in a new output table (`notices_pred.parquet`). This design enforces schema consistency between training and inference and supports reproducibility.

The deployment targets a decision support solution where the stakeholders can compare opportunities, assess expected contract values, and review competition context without having to go through multiple World Bank sources. Therefore, the prediction output is not treated as a standalone solution, it rather is complemented by a curated data layer that contains also all the other datasets and forms it into one single data layer. The Power BI dashboard loads the prepared tables (including `notices_pred.parquet`) and exposes a consolidated view that supports opportunity screening, predicted value-based prioritization, and competition-oriented analysis.

To ensure continuous availability of up-to-date insights, the end-to-end workflow is executed as a scheduled pipeline run. The data is getting then updated and predictions are applied with the pretrained model, which is excluded from being retrained. The retraining and reevaluation can be performed manually if change is necessary.

Deployment in this setting is subject to practical constraints typical for document-based public procurement data. First, data availability and disclosure delays can cause temporal sparsity or sudden changes in coverage. Second, heterogeneous document layouts can introduce extraction noise that influences features and predictions. Consequently, the deployed solution should be considered as a decision support rather than an automated process. Predictions are presented as estimates that should be interpreted together with contextual procurement information and domain judgment from the stakeholders. To support maintainability, the clear separation of pipeline steps enables iterative improvement of extraction rules, feature engineering, and model versions while keeping the stakeholder interface stable.

4 RESULTS

This chapter reports the empirical outcomes of the developed decision support solution. The results are structured according to predictive performance relative to simple baselines, diagnostic and robustness analyzes in value ranges and procurement segments, and interpretation and decision-support implications for stakeholder use.

4.1 Model Performance

The final supervised learning table used for award-value regression contains $N = 208{,}190$ award observations, split at the `ProjectID` level into $159{,}569$ training rows and $48{,}621$ test rows. The ProjectID-level split is crucial in this context because multiple awards share common project context which could lead to data leakage.

Table I summarizes the predictive performance on the test set. The CatBoost regressor is trained with a log-transformed target $\log(1 + y)$ and predictions are mapped back to USD scale via $\exp(\hat{y}) - 1$. The model is compared to two baselines, a global median predictor and a procurement category median predictor.

TABLE I: Test-set performance for award value prediction (USD scale). Baselines use training-only statistics.

Model	**MAE (USD)**	**RMSE (USD)**	R^2
CatBoost (log-target)	273,456	1,112,436	0.130
Global median baseline	297,955	1,221,269	-0.048
Category-median baseline	308,253	1,209,013	-0.027

Relative to the global-median baseline, the model reduces MAE by approximately 8.2% and RMSE by approximately 8.9%. Relative to the category-median baseline, MAE improves by approximately 11.3% and RMSE by approximately 8.0%. While the resulting R^2 remains low, the consistent improvement over the baseline medians indicates that the model captures systematic signal averages, despite big variance in the residuals which are typical for heterogeneous procurement data.

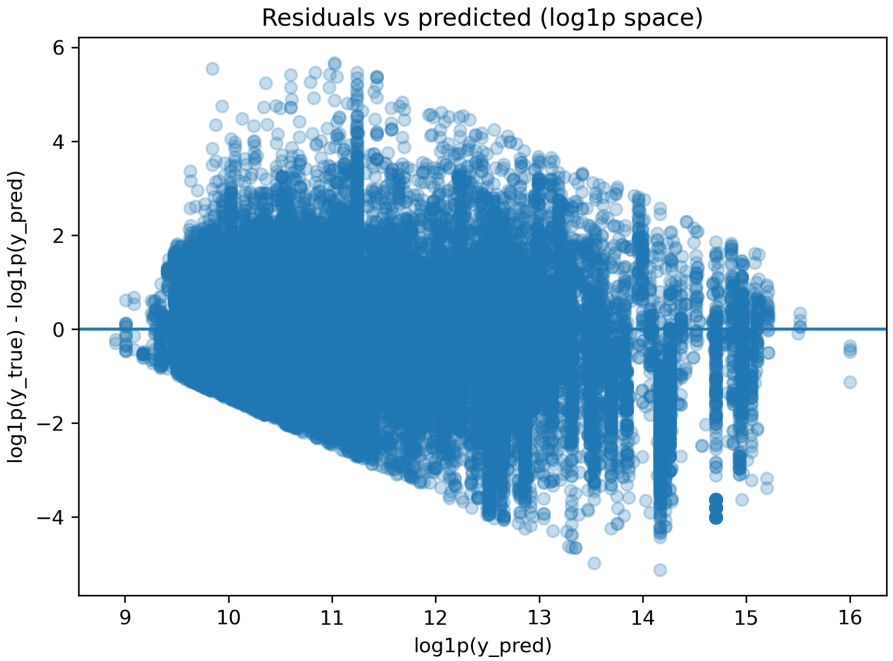

FIGURE 4: Residuals in $\log(1 + y)$ space vs. predicted values.

Figure 4 plots residuals in $\log(1 + y)$ space as a function of predicted values. The widening residual spread with increasing predictions indicates heteroscedasticity: higher-value awards are associated with larger absolute (and often also proportional) uncertainty. This supports positioning predictions as decision support estimates rather than point-accurate substitutes for pricing or budgeting.

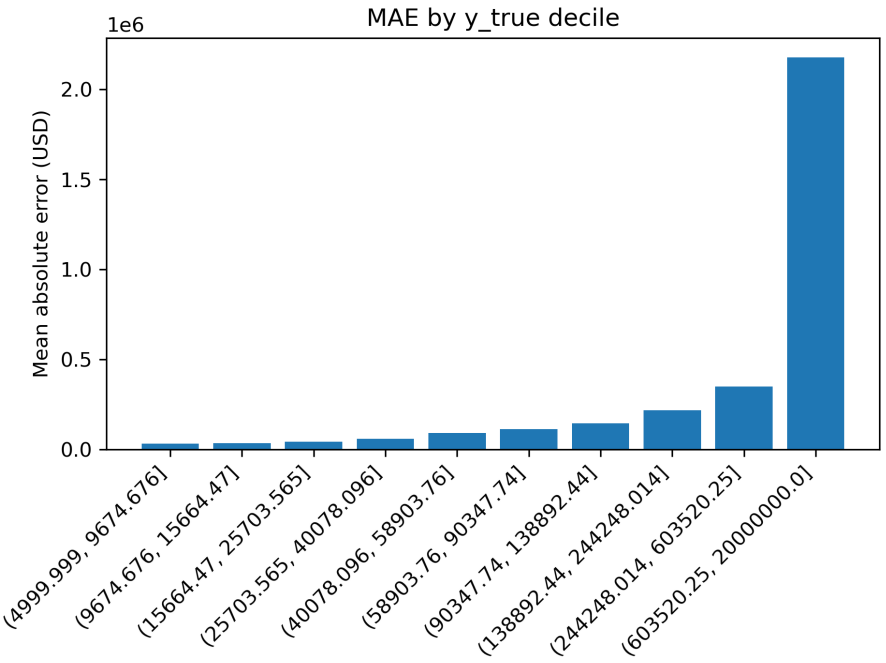

FIGURE 5: MAE by decile.

To assess whether errors concentrate in the high-value tail, MAE is analyzed across deciles of the true contract value distribution. Figure 5 shows a clear increase of MAE with contract size, with the largest decile (upper tail) contributing multi-million USD average absolute errors. This pattern is expected under heavy-tailed monetary outcomes and emphasizes that decision usefulness is highest when predicting low to medium value contracts, which are in the context of this specific stakeholder more interesting, rather than exact valuation of extreme cases.

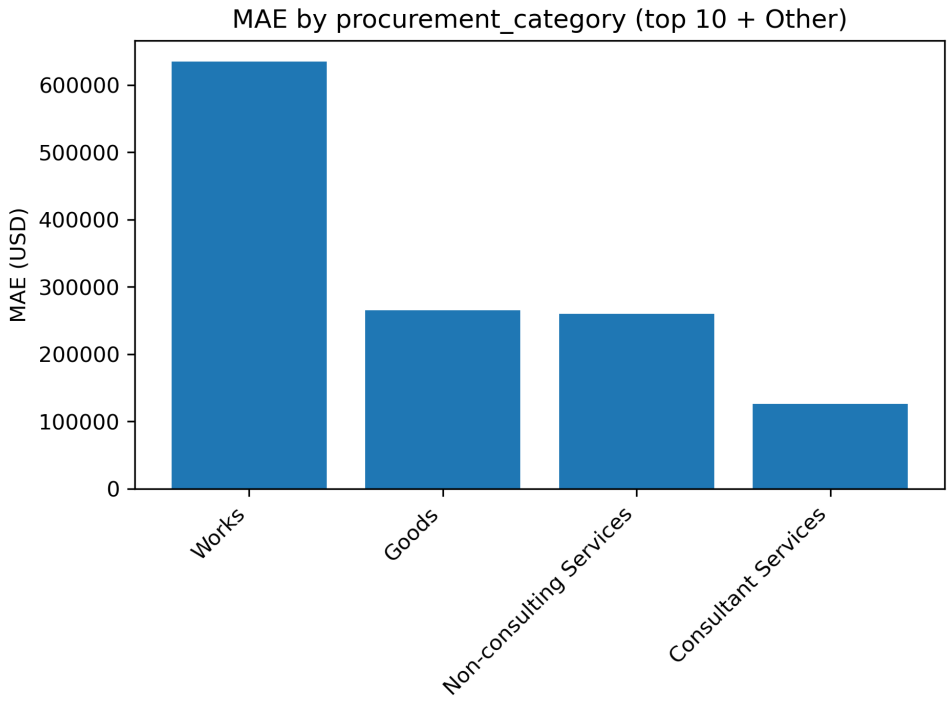

FIGURE 6: MAE (USD) by `procurement_category`.

Figure 6 reports MAE by `procurement_category`. As expected, categories associated with typically more high value contracts, like `Works`, show larger absolute errors than other categories. This is with domain knowledge expected since contracts in the category of `Works` are typically bigger contracts like infrastructure projects, which skew the data. For the case of the stakeholder who mostly compete in the `Consultant Services` and `Goods`, the MAE by category is far smaller indicating a stronger prediction performance for this specific areas.

Because the solution targets stakeholder-facing decision support, interpretability is treated as a core result. SHAP values are used to attribute predictions to individual features, enabling both global and case-based explanations.

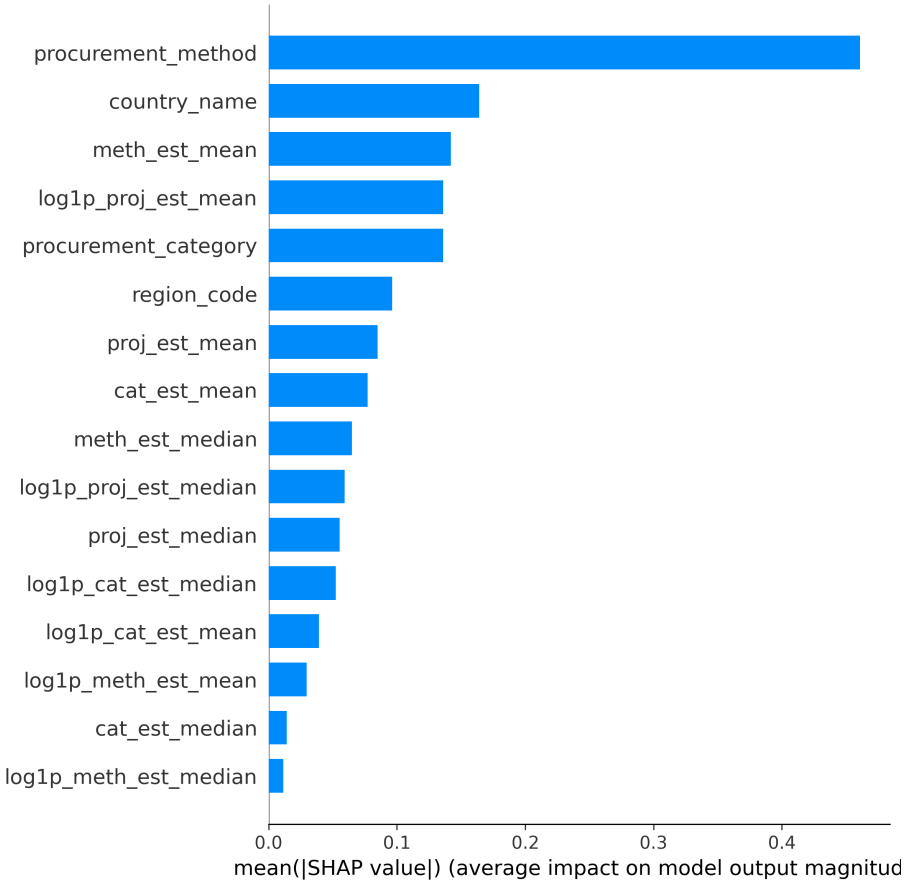

FIGURE 7: Global SHAP feature importance (mean absolute SHAP value).

Figure 7 ranks predictors by mean absolute SHAP value. The strongest drivers are categorical context variables (notably `procurement_method`, `country_name`, and `procurement_category`) and procurement-plan-derived monetary aggregates (e.g., method- and project-level estimated amount summaries). This pattern is consistent with the intended design of the feature space: categorical procedure context and planning-based magnitude signals jointly inform expected award values.

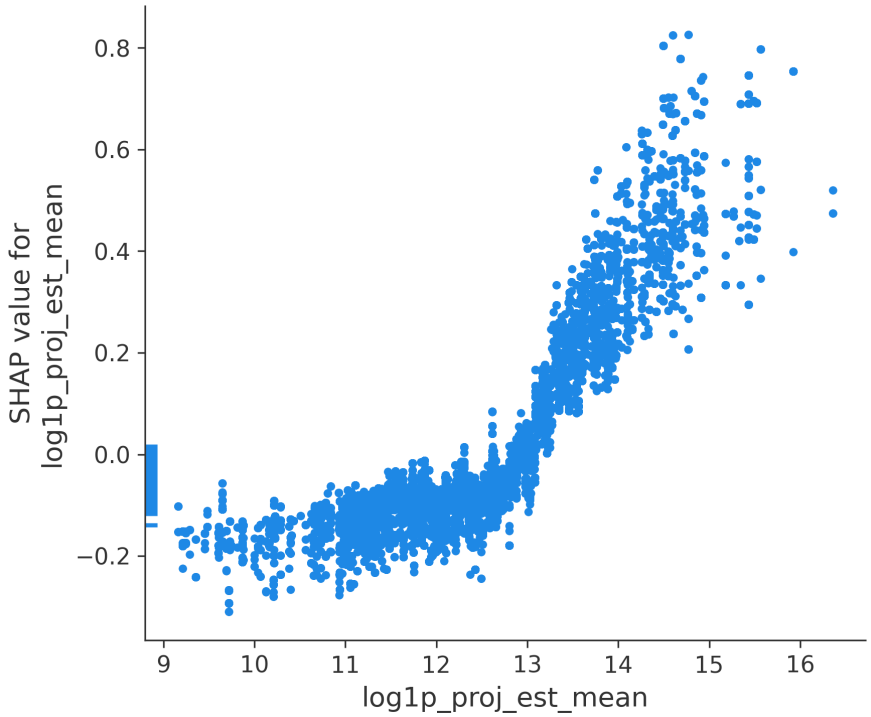

FIGURE 8: SHAP dependence plot for `log1p_proj_est_mean`.

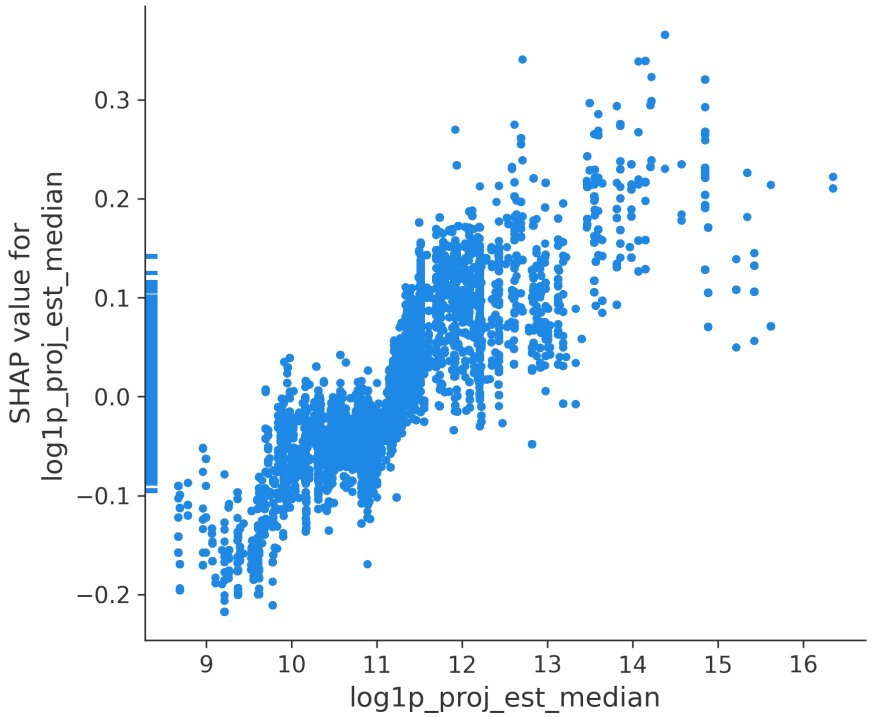

FIGURE 9: SHAP dependence plot for `log1p_proj_est_median`.

Figures 8 and 9 show dependence plots for log-transformed project-level estimated

amount aggregates. The overall trend indicates that higher estimated project amounts tend to increase the predicted award value, which aligns with domain expectations and shows that the model is learning economically meaningful structure rather than relying solely on categorical memorization. It also is expected that the importance of the median is higher for higher for lower estimated project amounts compared to the mean, given their statistical properties.

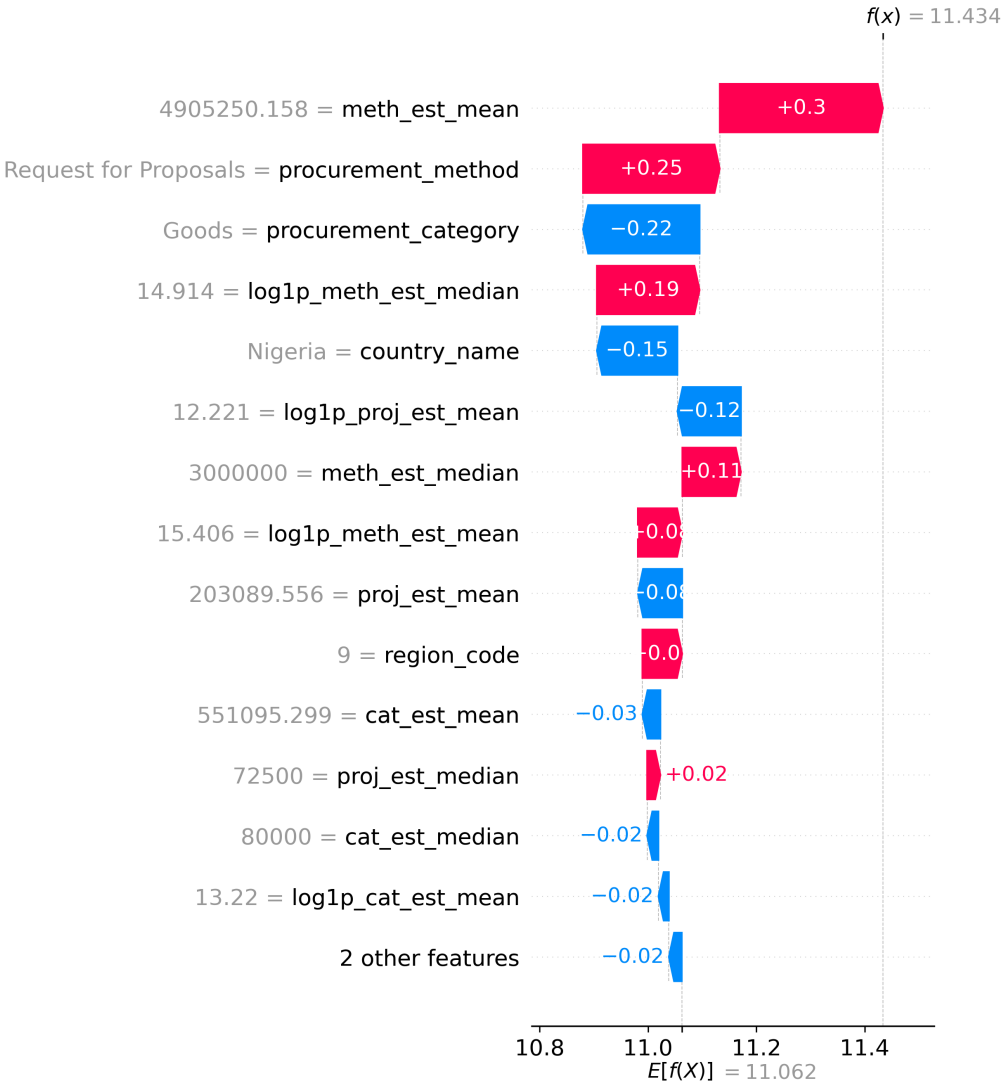

FIGURE 10: Local SHAP waterfall explanation for highest-error test instance (log space).

Figure 10 provides a waterfall explanation for the highest test error instance, decomposing the prediction into contributions relative to the expected value. Such case-based explanations support two stakeholder-relevant functions: transparency (why the model assigns a given estimate) and structured error analysis (which signals drove an over-/under-estimation). In practice, these explanations could be integrated into the dash-

board to stakeholders users interpret predictions. This was not yet done due to time issues but are a point of improvement for the future.

4.2 Objective-wise results

Objective 1: Unify heterogeneous procurement information into a single analytical layer. The results show that heterogeneous procurement artifacts can be consolidated into a curated analytical layer suitable for downstream analysis and BI consumption. The implemented pipeline transforms projects, procurement plans, notices, and awards into a unified Parquet layer. At the same time, the results also reveal a structural limitation of the source ecosystem. Linkage beyond `ProjectID` remains incomplete, such that a end-to-end link from a specific procurement plan line item to a specific award cannot be established in a reliable manner. As a consequence, the analytical layer supports artifact-level analysis well, while granular plan-to-award traceability remains constrained.

Objective 2: Provide early award-value estimates for upcoming opportunities. Model evaluation indicates that the CatBoost regressor captures predictive patterns beyond simple central-tendency heuristics. On the test set, it improves over transparent baselines in both MAE and RMSE. However, the results also show that absolute errors remain substantial in monetary terms, particularly for high-value awards, and the modest R^2 indicates that a large share of outcome variance is not explained by the available features. Taken together, the empirical evidence supports the use of the model as an early-stage screening signal, but not as a substitute for detailed pricing, budgeting, or bid preparation.

Objective 3: Provide trustworthy, interpretable outputs. Interpretability results demonstrate that explanation methods can be integrated into the prediction workflow in a technically robust manner. SHAP explanations provide both global and local feature attributions and indicate that model behavior is primarily driven by contextual variables (e.g., procurement method, geography) and derived monetary aggregates. Importantly, these results clarify why the model produces certain estimates, which supports transparency and reviewability. At the same time, the results underline that interpretability is not equivalent to correctness.

Objective 4: Reduce manual effort and support operational use via deployment. The deployment results show that the artifact can be used in a way that supports regular

use. Compute-intensive extraction, harmonization, and scoring are executed in a scheduled pipeline that produces refreshed analytical tables and predictions for BI consumption. This separation enables a stakeholder-facing dashboard that requires only a data refresh rather than repeated manual preprocessing. The operational usefulness is conditional on governance mechanisms around updates, versioning, and communication of limitations to prevent over-reliance on point estimates.

Taken together, the results suggest that the ML component provides measurable improvements over simple historical heuristics and yields explanations that are compatible with stakeholder interpretability requirements. At the same time, the remaining uncertainty particularly in the upper tail motivates utilizing model outputs as estimates for prioritization and screening, rather than as exact award-value forecasts.

5 DISCUSSION

The work performed in this project allowed satisfying the objectives initially proposed. The results show matching results to the literature, but are also deviating from the literature in some parts. The literature on procurement value prediction shows that ML can outperform simpler approaches, sometimes with substantial gains (García Rodríguez et al., 2021). In this work, improvements over baselines are real but not substantial.

Many published studies assume clean tender-to-award linkage at the transaction level and richer explanatory descriptors for each procurement. Here, linkage constraints required aggregation at joinable levels. Aggregations are informative, but they necessarily smooth away item specific details that can drive award values. This makes the task harder and limits achievable explanatory power. The usage of richer explanatory descriptors from NLP approaches to incorporate text features could be a big opportunity.

Plan-derived features depend on extracting numeric fields from PDF tables across heterogeneous layouts. Even with careful cleaning and schema alignment, extraction noise and inconsistencies can influence the signal, introduce missingness, and reduce effective feature quality. This issue is often underestimated, but it can dominate performance in real-world procurement ecosystems.

Procurement value is influenced by scope complexity, technical requirements, contract duration, financing conditions, and market dynamics. Many of these factors are not consistently available in structured form across World Bank artifacts or harder to retrieve. As a result, the model's unexplained variance remains high, which is reflected in low R^2. The limiting factor appears to be the feature space rather than the choice of algorithm.

5.1 Practical implications for procurement teams

For stakeholders, the central question is not whether the model is "good" in an abstract sense, but how it can change concrete decisions and workflows.

The primary contribution of the solution is faster and more structured opportunity screening. Predicted contract values can support an initial filtering of incoming notices, helping teams allocate attention and proposal resources to opportunities that appear financially meaningful and technically plausible given the companies capabilities. In addition, combining the predicted value with procurement-plan context and historical award patterns can reduce information asymmetry. Finally, when used alongside category and procurement-method filters, predictions can support planning by guiding decisions toward strategic focus areas where the company has domain expertise, competitive advantage, or empirically higher success rates.

Given the observed error structure and segment-level analyses, the model is most defensible as a decision-support signal for relative prioritization and early-stage screening across many opportunities. Its estimates are also suitable for rough sizing at the start of the pipeline, where the goal is to quickly identify which tenders are opportunities to compete and which are not. A further use is contextual comparison, in which predicted values are interpreted as a plausibility cue against project-level planning magnitudes rather than as a precise forecast.

The model should not be treated as exact pricing guidance. Absolute errorsespecially in the upper tailmake point predictions unsuitable as direct bid-price recommendations. It is likewise not a substitute for scope analysis, because the model cannot infer technical complexity, delivery constraints, or contract-specific risks that are central to budgeting. Finally, predictions should not be used as a stand-alone decision rule for "bid/no-bid" decisions. The output is a quantitative signal that must be interpreted within broader judgment. Operationally, the safest usage pattern is to include predictions into a structured decision checklist with explicit acknowledgment of uncertainty and escalation for high-value cases.

5.2 Risks

The observed dataset reflects only what is disclosed and captured through the available World Bank sources and extraction pipeline. Missingness and disclosure heterogeneity can induce selection effects (e.g., certain countries, methods, or project types being overrepresented). If the training data distribution differs from the opportunities the team is currently targeting, performance can degrade and the model may systematically misestimate values for underrepresented segments. Segment-wise error analysis already indicates heterogeneity, which should be treated as evidence that "one number fits all" is risky.

Document-based extraction can introduce systematic noise when extraction is not going as planned. This risk is partly mitigated by aggregation, but it cannot be eliminated. A practical mitigation is to implement automated data quality checks and to expose warnings in the dashboard when feature completeness is low.

Procurement format and disclosure practices change over time. Exchange rates, regulatory updates, and shifts in procurement methods can induce drift in both target scale and feature relationships. Without monitoring, a model trained on historical data can silently become less reliable.

Overall, the solution demonstrates that ML-supported procurement decision support is feasible and provides incremental value over simple heuristics, particularly for

early-stage estimation. However, the remaining uncertainty, concentrated in high-value cases and heterogeneous segments, requires careful framing, monitoring, and governance to ensure the system improves decisions rather than producing overconfident or biased guidance.

6 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

6.1 Conclusions

This report set out to improve decision making in World Bank procurement from a supplier perspective by reducing manual effort for information search, reducing information asymmetry, and supporting earlier, more structured opportunity screening through an end-to-end decision support artifact. The artifact consolidates heterogeneous procurement artifacts into a unified analytical layer and augments this layer with machine-learning-based award value estimation, interpretable outputs, and operational deployment.

6.2 Answer to the research question and objectives

The guiding research question asked whether machine learning can be effective for procurement decision making and thereby support more successful participation. The results support the effectiveness claim in a methodological sense. The award-value model improves over transparent heuristic baselines on standard regression error metrics, indicating that ML can capture additional signal beyond central-tendency rules, even under heterogeneous disclosure conditions. However, the second part "more success" cannot be established within the scope of this study because bid outcomes and organizational performance effects are not observed. Evaluating the effect would take a longer field study. Consequently, the contribution is best interpreted as enabling infrastructure and evidence for improved screening decisions rather than a measured increase in contracts won.

Across the objectives, the evidence is mixed but directionally positive. First, unifying heterogeneous procurement information into a single analytical layer is largely achieved: projects, procurement plans, notices, and awards are consolidated into curated analytical tables, and procurement-plan PDF tables are converted into a schema-consistent dataset. Second, providing early award-value estimates is partially achieved: while the model improves over baselines, absolute errors remain substantial in monetary termsespecially for high-value awardsand a modest R^2 indicates that a large share of variance remains unexplained. This supports using predictions for early-stage screening and ballpark sizing, but not as a substitute for pricing or budgeting. Third, providing trustworthy, interpretable outputs is achieved in a technical sense through SHAP-based explanations. Fourth, reducing manual effort through deployment is largely achieved through a pipeline-and-dashboard design consistent with end-to-end applied analytics practice, though sustained usefulness depends on governance and communication of limitations.

6.3 Why the artifact works despite a difficult data environment

Several design decisions, motivated by the literature and by procurement data characteristics, are supported by the results and help explain why the artifact yields measurable improvements despite difficult data conditions.

Splitting train/test data at the `ProjectID` level reduces the risk of overly optimistic evaluation caused by correlated observations within the same project context. While this tends to lower headline performance metrics, it increases credibility because it aligns evaluation with the intended deployment setting of generalizing to previously unseen projects. In procurement analytics, this is an important safeguard against hidden leakage when artifacts share contextual identifiers and repeated structures across records.

Procurement award values are heavy-tailed, so diagnostic plots in log space are informative: they show that the model captures broad variation while struggling in the extreme upper tail. Modeling the target on a log scale is therefore justified because it stabilizes learning under heavy-tailed outcomes and encourages proportional error minimization. Practically, this supports the intended role of the model as an early-stage estimator rather than an exact forecaster.

Comparisons against global and category medians provide an operational relevance check because stakeholders often apply informal heuristics based on typical contract sizes. Improvement over these heuristics indicates added value, while the remaining gap shows the reality, that residual uncertainty in procurement cannot be eliminated by tabular context alone, particularly when key scope drivers are not observable.

6.4 Main findings and contributions

The work contributes an applied end-to-end template for procurement decision support under document-based disclosure. At the data level, it demonstrates that heterogeneous procurement artifacts aligned with the procurement lifecycle can be consolidated into an analytical layer grounded in standard process and artifact definitions. At the modeling level, it provides evidence that gradient-boosted decision trees designed for categorical data (CatBoost) can outperform transparent baselines for award value estimation under realistic constraints. At the accountability level, it integrates model interpretability via SHAP to support stakeholder-facing explanation and structured error analysis. At the operational level, it treats deployment as a first-class phase in line with CRISP-DMs end-to-end framing.

6.5 Limitations

6.5.1 Data limitations

A core dependency of the feature set is the extraction of estimated amounts from procurement-plan PDF tables. PDF layouts vary across projects and countries, include multi-row headers and repeated header blocks, and may contain pages that are difficult to extract information from. Even with careful cleaning and schema alignment, residual extraction errors can compromise predictive signal or introduce systematic bias. This limitation is structural to document-based procurement disclosure and is likely a dominant contributor to unexplained variance.

While `ProjectID` enables integration at the project level, end-to-end linkage from procurement plan row items to specific notices and awards is not possible. As a consequence, feature engineering relies on aggregation at joinable levels (project/method/category), which necessarily discards item-specific scope information. This constrains the maximum achievable precision of award value estimates.

6.5.2 Method limitations

Despite ProjectID-level splitting, leakage risk cannot be eliminated entirely in observational procurement data. Aggregations derived from procurement plans and historical awards may implicitly encode information correlated with the target in ways that differ between training and deployment. For example, if some projects contain plan updates that reflect realized values, plan-derived aggregates could become partially influenced. The implemented split strategy reduces this risk, but it remains a methodological limitation inherent to real-world procurement disclosures.

Award value is strongly influenced by scope complexity, contract duration, technical requirements, financing conditions, and market dynamics. These drivers are not consistently captured in the structured datasets used in this work. As a result, the model may rely on proxy variables that are only partially informative. This is not a failure of the algorithm but a limitation of the information available at prediction time.

The artifact outputs a point estimate. Given the heavy-tailed distribution and heteroscedastic residual structure, stakeholders would benefit from uncertainty-aware outputs (prediction intervals, quantile estimates). Without explicit uncertainty, there is a higher risk of overconfidence and misuse, especially for high-value opportunities where absolute errors are largest.

6.6 Future work

6.6.1 Short-term

A practical next step is to make outputs uncertainty-aware by extending point predictions to quantile regression or bootstrapped prediction intervals. In parallel, stronger automated data-quality validation should be added, including plausibility checks for extracted values and consistency checks between plan-derived estimates and award magnitudes. The resulting quality flags should be persisted and surfaced in Power BI. Evaluation can be made more operational by institutionalizing segment-aware monitoring to detect drift early. Within existing sources, additional feature enrichment is feasible through project sector/theme, temporal variables (year/quarter), and plan-structure signals such as counts and dispersion measures. Finally, stronger baselines would improve the robustness of performance interpretation.

6.6.2 Long-term

In the longer term, richer scope representations are likely to yield the largest gains. Incorporating text-derived features from notice and plan descriptions via NLP could capture technical complexity that is currently omitted. Improved linkage and entity resolution across the lifecycle via a more elaborate record linkage effort would enable more precise, item-level feature engineering. As the system matures, procurement-appropriate governance becomes essential, including a model registry, versioned data snapshots, change logs, and formal usage policies to ensure auditability when outputs influence resource allocation decisions.

Overall, this work shows that meaningful procurement decision support can be implemented under realistic constraints of heterogeneous, document-based disclosures. The artifact translates fragmented World Bank procurement information into an operational analytical layer and provides an interpretable early-stage estimate that improves on transparent heuristics, while clearly delineating where uncertainty remainsespecially for high-value awards. The outlined limitations and future directions therefore constitute a concrete roadmap for increasing robustness, scope sensitivity, and governance, moving the system from a thesis artifact toward a sustainable tool for supplier-side opportunity screening.

REFERENCES

Adhikari, N. S. and Agarwal, S. (2025), 'A Comparative Study of PDF Parsing Tools Across Diverse Document Categories'. DOI: 10.48550/arXiv.2410.09871.

Botchkarev, A. (2019), 'A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms', *Interdisciplinary Journal of Information, Knowledge, and Management* **14**, 045076. DOI: 10.28945/4184.

Chapman, P. (2000), CRISP-DM 1.0: Step-by-step data mining guide.
URL: *https://api.semanticscholar.org/CorpusID:59777418*

Costa, C. J. and Aparicio, J. T. (2020), POST-DS: A Methodology to Boost Data Science, *in* '2020 15th Iberian Conference on Information Systems and Technologies (CISTI)', pp. 1–6. DOI: 10.23919/CISTI49556.2020.9140932.

Csáki, C. and Prier, E. (2018), *Quality Issues of Public Procurement Open Data: 7th International Conference, EGOVIS 2018, Regensburg, Germany, September 3, 2018, Proceedings*, pp. 177–191. DOI: 10.1007/978-3-319-98349-$3_1$4.

Fazekas, M. and Kocsis, G. (2020), 'Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data', *British Journal of Political Science* **50**(1), 155164. DOI: 10.1017/S0007123417000461.

Fazekas, M., Tóth, B., Abdou, A. and Al-Shaibani, A. (2024), 'Global Contract-level Public Procurement Dataset', *Data in Brief* **54**, 110412. DOI: 10.1016/j.dib.2024.110412.

García Rodríguez, M. J., Montequín, V., Aranguren Ubierna, A., Santana, R., Sierra, B. and Jauregi, A. (2021), 'Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain', *Studies in Informatics and Control* **30**, 67–76. DOI: 10.24846/v30i4y202106.

García Rodríguez, M. J., Montequín, V., Ortega-Fernández, F. and Balsera, J. (2019), 'Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning', *Complexity* **2019**, 1–20. DOI: 10.1155/2019/2360610.

Hodson, T. O. (2022), 'Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not', *Geoscientific Model Development* **15**(14), 5481–5487. DOI: 10.5194/gmd-15-5481-2022.

Jensen, K. (2012), 'CRISP-DM Process Diagram'. Accessed: 2026-01-24. License: CC BY-SA 3.0.
URL: *https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png*

Kim, J.-M. and Jung, H. (2018), 'Predicting bid prices by using machine learning methods', *Applied Economics* **51**, 1–8. DOI: 10.1080/00036846.2018.1537477.

Lundberg, S. and Lee, S.-I. (2017), A Unified Approach to Interpreting Model Predictions. DOI: 10.48550/arXiv.1705.07874.

Nai, R., Meo, R., Morina, G. and Pasteris, P. (2023), 'Public tenders, complaints, machine learning and recommender systems: a case study in public administration', *Computer Law Security Review* **51**, 105887. DOI: 10.1016/j.clsr.2023.105887.

Obwegeser, N. and Müller, S. D. (2018), 'Innovation and public procurement: Terminology, concepts, and applications', *Technovation* **74-75**, 1–17. DOI: 10.1016/j.technovation.2018.02.015.

OpenAI (2026), 'AI-generated image: Procurement artifact hierarchy pyramid', Generated with DALL-E via ChatGPT. Prompt: "Generate a pyramid kind of like the food pyramid but I want the hierarchy of Project, Procurement plans, notices and award contracts. Make the image in this order. Make it not too colorful and minimal so it fits in a academic paper". Generated: 2026-01-26.
URL: *https://chatgpt.com/s/m_69778b38f9848191a1db61f07ae18209*

Open Contracting Data Standard (n.d.), Online documentation.
URL: *https://standard.open-contracting.org/latest/en/*

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. and Gulin, A. (2019), 'Catboost: unbiased boosting with categorical features'. DOI: 10.48550/arXiv.1706.09516.

Schröer, C., Kruse, F. and Gómez, J. M. (2021), 'A Systematic Literature Review on Applying CRISP-DM Process Model', *Procedia Computer Science* **181**, 526–534. CENTERIS 2020 - International Conference on Enterprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020. DOI: 10.1016/j.procs.2021.01.199.

vom Brocke, J., Hevner, A. and Maedche, A. (2020), *Introduction to Design Science Research*, Springer International Publishing, Cham, pp. 1–13.
URL: *https://doi.org/10.1007/978-3-030-46781-4*$_1$

Wirth, R. and Hipp, J. (2000), 'CRISP-DM: Towards a standard process model for data mining', *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* .
URL: *https://www.researchgate.net/publication/239585378_CRISP–DM_Towards_a_standard_process_model_for_data_mining*

World Bank (2025), *The World Bank Procurement Regulations for IPF Borrowers: Procurement in Investment Project Financing - Goods, Works, Non-Consulting and Consulting Services - Sixth Edition*, sixth edn, World Bank Group, Washington, D.C. Operational Procurement Guidance and Form, Document Number 197185, Final Version.
URL: *http://documents.worldbank.org/curated/en/099120102072534901*