Practical Applications of Machine Learning and Artificial Intelligence in Predicting the Cost of Consulting Projects

Abiodun Gabriel Ajanaku
Independent Machine Learning (ML) and Artificial Intelligence (AI) Researcher

Abstract

Background: In the consulting economy, accurate prediction of project costs is a significant challenge due to several factors including but not limited to diverse scope in terms of project complexity, duration, resources requirements; limited historical data, and negotiation pressure leading to lower cost estimates. Consulting projects are tailored to client-specific needs, making it difficult to generalise across projects. The consulting industry is highly dynamic and project cost estimations often depends on complex non-linear variables that are difficult to standardize or predict. Traditional cost estimation methods have relied heavily on historical benchmarking and expert judgement; overlooking non-linear interactions amongst critical factors that influence project costs. This study is aimed at: (1) Identifying the critical factors in predicting the cost of consulting projects using machine learning and artificial intelligence techniques, (2) Identify, compare the performance of 5 machine learning models, and select the most suitable model for building a consulting project cost predictive system, (3) synthesise existing knowledge to identify gaps to support the training, validation, testing and deployment of a consulting projects cost predictive model.

Method: A quantitative analysis was carried out through a comprehensive examination of 80 real-world diverse consulting projects executed over the last 12 years by a consulting firm. Domain expertise knowledge was leveraged in identifying and prioritising the critical factors influencing consulting projects cost during the investigation. The identified factors were categorized into project-specific, client-specific, and industry-relevant financial data. A combination of rule-based and simulation techniques using programming and structured data science techniques were deployed in simulating the real-world project scenarios to generate 2000 synthetic project dataset with 17 features. The dataset was used to construct five machine learning models including Linear Regression (LR), Decision Tree Regressor (DT), Random Forest Regressor (RF), XGBoost (Extreme Gradient Boosting)**,** and Support Vector Machine (SVM) to predict the cost of consulting projects. The performance of the models was evaluated using success metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Squared Error (RMSR) and R^2 to gauge how the model penalised and minimize prediction errors. Model optimization was carried out through cross validation and hyperparameters tuning. Random search was deployed to select the best hyperparameters and establish how the model perform and generalise on unseen test dataset.

Findings*:* The XGBoost model had the best performance with MAPE of 5%, MAE of 1.041e+05 and R^2 of 99%. Hence, outperforming the other models in terms of minimisation of prediction errors. Furthermore, the results from Mutual Information (MI), Random Forest Regressor, and SHAP feature importance all indicated that person-months ('level of human effort') devoted to a project emerged as the strongest explanatory feature with (MI score of 89.7%) and holds significant information about the total project cost ('target variable'). Hence, when combined with other relevant features such as team size, project duration, frequency of travels, and other expenses; it helps to correctly map the complexity of any project as either LOW, MEDIUM, or HIGH.

Conclusion: The study showed that machine learning can accurately predict the cost of consulting projects using baseline information such as person-months, person-month (PM) rate, project team size, duration, frequency of travel and other direct human and materials resources. These relevant factors are what define the complexity of any consulting project. Project with medium to high complexity requires a moderate to high person-months. For consulting firms, the ability to accurately estimate project cost is not a financial imperative but a pathway to building trust with clients and achieving a competitive advantage. Hence, further research is necessary to integrate real-time changes in project scope, client satisfaction, economic and market data to enhance ML model performance, generalisation, and more dynamic cost predictive model.

Keywords: Artificial Intelligence, Machine Learning, Consulting Projects, Cost Estimation, Management Consulting.

Abbreviations: AI, Artificial Intelligence; ML, Machine Learning; COCOMO, Constructive Cost Model; MOGA, Multi-Objective Genetic Algorithm; MARE, Mean Absolute Relative Error, MMRE, Mean Magnitude of Relative Error (MMRE); CRISP-DM, Cross Industry Standard Process for Data Mining Model

* Corresponding author: Abiodun Ajanaku | E-mail address*:* **biodunajan@gmail.com**

1. Introduction

The wave of technology is moving at an accelerating rate. The technology that comes with **Machine Learning (ML) and Artificial Intelligence (AI)** is currently replicating and at the same complementing human expertise in accurately estimating project costs and other professional services. In today's competitive consulting landscape, cost estimation is a critical component of project planning and accurate prediction of consulting projects is important for both the service providers and the clients for many reasons including ***resources allocation, risk management, capital optimisation, maximization of desired impacts for all stakeholders amongst others***. Over the years, consulting firms have relied on traditional methods in forecasting the financial proposal of projects; these methods range from domain expert judgement, historical data, comparative projects executed in the past, industry standards amongst others. Despite the several merits with these methods, it is heavily prone to human biases, subjective decision making, inaccurate predictions (under/over budget), difficulties in scaling, incomplete information and do not accurately capture the inherent complexities and nuances of consulting projects. As consulting projects grow in scope, diversity, and intricacy; the traditional methods are now very weak in producing a data-driven and more dynamic project cost estimate with a high level of accuracy to drive consulting engagements and decision-making process.

The applications of ML and AI offer massive opportunity to address the limitations associated with the traditional methods of estimating the cost of consulting projects. ML and AI models are effective predictive algorithms because they can identify linear, complex non-linear relationships and can learn from big historical data to make objective and repeatable predictions. While existing research and papers have applied ML and AI to cost estimations in other industries such as construction, information technology (IT, software), there is limited studies specifically focused on dynamic industry such as consulting: consulting projects are unique due to their highly variable nature, reliance on human expertise, and subjective factors such as client relationship and importance of project.

1.1 Literature Review

In recapping related works on the research topic, the Social Science Research Network (SSRN) database was searched from inception to December 31, 2024 using the search string: ***"Artificial Intelligence" OR "AI" AND "machine learning" OR "ML" AND "consulting projects" OR "project cost estimation" OR "management consulting"***. The search returned 10,000 unique papers of which 9,900 articles were excluded because they did not meet the inclusion criteria, were duplicates or were related to other studies. Full-text reviews were conducted for 100 articles and a final set of 6 articles were included for literature review because they deployed AI/ML techniques for cost estimations and automation in different sectors such as consulting, construction, I.T, and related domains. The study in [1] explored the methodology of software cost estimation using machine learning techniques. The research applied genetic programming, neural networks models and testing was conducted using COCOMO 81 data set. They applied fuzzy C-means clustering algorithm for project clustering. Then, parameters of COCOMO model have been optimized using Multi-objective Genetic Algorithm. The result has proved superiority of MOGA in parameter optimization for getting strength back the accuracy of software cost estimation. The performance of the models was investigated using parameters such as the MARE and MMRE. The experimental outcomes have demonstrated that ML methods outperforms the existing method in estimating software effort more precisely. The authors in [2] carried out an applied algorithmic machine learning for intelligent project prediction towards an AI Framework of Project Success. The study utilized an empirical investigation using ethnographic practice-based exploration at Company Alpha to establish overlapping features of projects within the same portfolio with the aim of capturing the emergent inter-project relationships. The research compared two specific types of artificial neural network prediction performances; (i) multilayer perceptron and; (ii) recurrent neural networks. The research validated that the multilayer perceptron has been found to be one of the most widely used artificial neural networks in the project management literature, and recurrent networks are distinguished by the memory they take from prior inputs to influence input and output. The findings of the study confirmed that recurrent neural networks can capture the potential inter-project relationship towards achieving improved performance in contrast to multilayer perceptron.

Gao et al. explored a data-driven machine-learning-based analytic pipeline to predict, interpret, and infer the nature and severity of cost or schedule overrun a project will experience. The study built a novel metric to identify and interpret at a granular level the top predictors that contribute to cost or schedule overrun and the severity of each variable. These variables include contract pricing, contract types, incentives amongst others. The study confirmed the influence of contract pricing on project performance but also find that existing project incentives are lacking in effectiveness in managing cost and schedule overruns. In addition, the study validated that cost and performance incentives, when combined with fixed-price contracts, seem to cause additional cost and schedule overruns. However, performance incentives can mitigate cost overrun when offered with cost-plus contracts[3]. In the study carried out by Dzhusupova et al, they examined an innovative Artificial Intelligence (AI) - based solution for enhancing material procurement and cost estimation in Engineering, Procurement and Construction (EPC) projects, particularly within the oil, gas, and petrochemical sectors[4]. The work utilized deep learning techniques for symbol and text recognition with regression analysis of historical project data with the of automating the traditionally manual and labour-intensive process of estimating material quantities. The findings of the research emphasized that machine learning techniques offer significant benefits material procurement and cost estimation in EPC which include time savings, reduction in human error, and cost-effectiveness, thereby facilitating sustainable and competitive EPC project bids. As shown in [5], the applications of machine learning in financial accounting with a comprehensive case study on cost estimation and forecasting. The study utilized historical financial data and operational metrics various supervised and unsupervised learning algorithms, including regression models, decision trees, and neural networks, were deployed to enhance the accuracy of cost predictions and streamline financial decision-making processes. The results of the study highlighted the significant advantage and benefits of machine learning techniques such as precision of cost estimations and identification of patterns and anomalies that are difficult to be spotted using traditional accounting methods. Kan et al. deployed interpretable machine learning techniques for cost prediction for water reuse equipment. In the study, they deployed four traditional models, and four ensemble models to predict the cost of water reuse equipment. Features such as water quantity, inflow conductivity, outflow conductivity, and recovery rate emerged as critical factors affecting the cost of water reuse equipment. The results of the study showed that the ensemble models exhibited significantly superior predictive performance to that of traditional models, with the three boosting ensemble models achieving the best performance (training set R2 = 97.42%, testing set R2 = 82.16%-93.79%). The techniques utilized in the research demonstrated that ML could enhance the versatility of cost prediction processes in environmental engineering scenarios, particularly those concerning the construction costs of water treatment equipment[6].

1.2 The Applications of Machine Learning and Artificial Intelligence

AI is any program that can replicate human intelligence: sense, reason, adapt and act on any intelligence task the same way and at much better level of accuracy than humans. Alan Turing, as cited in [7] posed the question: “Can machines think?” in his paper “Computing Machinery and Intelligence” [7]. AI replicates human intelligence and it is self-learning in detecting patterns in big data by clustering information that can be deployed in identifying relevant features. It is the branch of computer science that aims to answer the question posed by Alan Turing. There are two general classifications of AI: the narrow AI and the artificial general intelligence (AGI). The narrow AI only works in a limited context and focus on performing one task at a time. Whereas the AGI is the future of AI and have general intelligence that could be applied to solve any problem[7].

ML is a core branch of AI and extension of traditional statistical techniques and mathematics that uses computational resources. The ML algorithms learn progressively through continued exposure to data and refinement of other results that iteratively improve the machine's ability to make decisions, detect underlying patterns in high-dimensional data, and it is increasingly being used in different areas in healthcare, finance and other domains requiring predictions. The goal of machine learning models/algorithms is to use a building block approach to facilitate and establish the best way in mapping of features (inputs) to output (labels) through learning from the dataset. In other words, ML identify relevant factors, classify, and organize different types of information to predict or classify something. The task of classify/labelling data into multiple categorical classes is called classification. On the other hand, the process of making a numeric prediction based on independent variables is called regression. Both the classification and regression problem-solving approach are categorised under the supervised machine learning [8].

The other types of machine learning are unsupervised machine learning: this is data driven ML techniques used for clustering or identifying clusters in a data set. The third type of machine learning is the reinforcement learning which allows a machine to learn from mistakes; overtime reinforcement models learn to make optimal choices. The application of these three forms of training helps computers to learn to solve practical problems such as predicting a disease breakout, fraud detections, image classification and making other real-time decisions[7]. AI and ML programs performance improve as they are exposed to big data over time. In addition, they provide swifter turnaround times in reviewing documentation, premised on the basis that AI can be utilised to identify relevant features for ML predictions and other relevant information quicker than humans. Hence, the applications of ML and AL techniques in predicting the cost of projects will deliver the benefits and revolutionize how consulting firms project pricing, resource allocations, and overall project management and execution.

The five models selected for this research project includes Linear Regression, Decision Tree Regressor, Support Vector Machines (SVMs), Random Forest Regressor and XGBoost (eXtreme Gradient Boosting). They have been selected to examine the linear interactions amongst the features (Linear Regression); provide insights into the features with the most important features for predictions, robustness to handle outliers and less sensitivity to overfitting (Random Forest Regressor); handle high-dimensional datasets with linear and non-linear complex relationships amongst the variables (XGBoost).

1.3 The Significance of the Study

The aim of this research paper is to explore the applications of ML and AI in predicting the cost of consulting projects by investigating and integrating a wide range of features that reflect the complexities and diverse factors that influence consulting fees. These factors include project-specific data, client-specific data, and industry-relevant financial data. The outcome of the research is to develop and deploy ML/AI model to predict consulting projects and enhance decision-making using relevant factors identified in the research. This model will provide consulting firms with an invaluable tool that support decision-making, better resource allocation and budget tracking, ultimately leading to profitable and efficient project implementation.

The insights and outcomes of the study will provide significant benefits to a wide range of stakeholders across different roles and industries as detailed in Table 1 below. The research combines data science techniques with industry specific knowledge, the goal is to address the challenges faced by consulting firms in predicting project costs and demonstrate the significant potential of ML/AI in transforming consulting practices.

Table 1: Primary and Secondary Beneficiaries of the ML/AI Consulting Projects Cost Predictive Model

Category	Beneficiaries	Benefits
A. **Primary Beneficiaries**: These beneficiaries directly engage with or depend on the cost estimation processes	1. Consulting firms	• **Accurate Pricing for Financial Proposals**: the ML/AI predictive model will serve a decision support system for generating financial proposals and pricing with a level of accuracy require for competitive and realistic pricing proposals. • **Other benefits to consulting firms include but not limited to** improved profitability, resource allocation, risk management, reduction of overhead cost through automation and scalability.
	2. Clients (Organisations Hiring Consultants)	• **Competitive Tender/Bidding Process:** Organisations hiring consultants will benefit from transparent pricing. Hence, reducing the risks of overcharge or hidden costs. • **Other benefits to clients include but not limited to** better budgeting and financial planning, better negotiation process, increased trust through data-driven decision-making. Hence, minimizing subjective pricing.
	3. Project Managers and Consultants	• **Project Budget Planning and Time Saving.** Project Managers will utilize the ML/AI model for efficient resource allocation and budget planning. This will also reduce the time spent on manually developing project cost estimation. • **Other benefits to Project Managers/Consultants include but not limited to:** Data-driven decision making and forecasting future project with a high level of accuracy.
	4. Financial Teams	• **Better Project Budget Tracking and Controls:** The finance teams can use the model for monitor actual and predicted expenses for cost controls and financial governance. • **Other benefits to finance teams include but not limited to** better revenue forecasting, accurate reporting to project funders and financial analytics to uncover patterns and insights from project budget.
B. **Secondary Beneficiaries**: These beneficiaries indirectly benefit from improve cost prediction practices	5. Executives and Business Development Teams	• Strategic decision making to establish profitable projects, market opportunities and diversification.
	6. Industry Analysts and Researchers	• Aggregated data from the models can be utilized as a pipeline for benchmarking projects across industries and regions.

	7. Technology and Operations Team	• The ML/AI model helps automates routine tasks relating to cost estimations. Hence, reducing the workloads operational staff. In addition, the model can be integrated into other enterprise workflow to improve operational efficiency.
	8. Regulatory and Compliance Authorities	• Regulatory and Compliance Authorities can utilize the model for standardization of pricing practices across consulting industry, benefiting compliance and regulatory efforts.
	9. Academic Institutions	• The ML/AI model can be utilized as use cases to educate students about the applications of AI/ML in consulting domains and cost estimations
	10. Entrepreneurs and Startups in Consulting	• Entrepreneurs and startups in consulting services without extensive historical data or experience can use the model to accurately predict the cost of new client proposals

2. Methodology

The study employed a quantitative approach to identify and analyse the relationships amongst several factors influencing consulting projects cost with the aim of developing a project cost predictive model using AI/ML techniques[9]. The study explored the application of AI/ML to real-world consulting projects by investigating 80 industry-relevant consulting projects executed over the last 12 years by a consulting firm. In addition, it leveraged domain knowledge to identify relevant factors; simulating these features to generate 2000 project scenarios/dataset for model training, validation, and testing.

2.1. Study Design & Sampling: Investigating of Real-World Consulting Projects

To ground the research in practical relevance, the below outlined the study design and sampling process.

- **Project Selection:** Industry-relevant projects that cut across relevant sectors such as **education, start-up-supports, health, investments, fintech, creative economy, research and development, digital security amongst others** were investigated to ensure coverage of varied business context and challenges.
- **Data Collection:** Information was gathered through a combination of project documentation reviews, finance experts' reviews, and analysis of the various financial element of financial proposals.
- **Key Insights:** From the analysis, 17 key factors influencing consulting projects were identified as detailed in Table 2 below:

Table 2: Identified Features from the investigation of 80 real-world diverse consulting projects.

S/N	Features/Factors	Variables	Meaning and Unit of Measurement	Data Type
1.	Project Complexity	complexity	The complexity of a project is defined by its scope, number of deliverables/ work packages, level of risks and other resources or requirements. It is categorized into LOW (1), MEDIUM (2) and HIGH(3).	String/Text
2.	Team Size	team_size	Team size is measured by no of experts or personnel devoted to a project	Integer
3.	Duration	duration	Project duration measures in months: ranging from as low as 1 to as high as 120 months.	Integer
4.	Person-Months (PM)	person_months	Person-months (also called man-months or 'human effort') is a unit of measure used in to estimate the amount of effort devoted to completing a task or project [10]. It represents the amount of work that one person can complete in one month, assuming full-time work (typically 40 hours per week).	Integer
5.	Average Person-Months (PM) Rate	person_month_rate	The average PM rate represents the cost of one person working full-time for one month on a project [11].	Integer
6.	Industry	industry	These are sectors where consulting project have been executed this include the below with their respective assigned label encoding: (1=Education, 2=Healthcare, 3=Creative Economy, 4=IT, 5=Startup Supports,6=Public	String/Text

			Sector, 7=Agritech, 8=Fintech, 9=Research, 10=Green Tech, 11=Cybersecurity,12=Others)	
7.	Project Type	project_type	This is the classification of the project based on their objectives, nature, scope, and industry as listed: 1=Research & Development, 2=Advisory, 3=Training & Capacity Building,4=Grant Management, 5=Audit services, 6=Project Management, 7=Innovation Consultany8=Fund Management, 9=Others).	String/Text
8.	Consultant/Technical Advisor Level	consultant_level	This is the ranking of technical advisor and consultant on consulting project based on the level experience, expertise, responsibilities, and role in delivering consulting services. (1=Junior, 2=Mid-level, 3=Senior)	String/Text
9.	Frequency of Travel	frequency_of_travel	This is the number of trips/travels on a project. High, medium, and low complexity project requires high, moderate, and relatively low frequency of travels respectively.	Integer
10.	Number of Travellers	num_travellers	This is the number of project staff travelling per trip on a project. High, medium, and low complexity project requires high, moderate, and relatively low people travelling respectively.	Integer
11.	Average Travel Days	avg_travel_days	This is the number of days per day per trip on a project.	Integer
12.	Total Travel Cost	total_travel_cost	This is the total travel cost for a project. It is the combination of average travel cost, number of persons travelling and the average travel days.	Integer
13.	Average Event Cost	avg_event_cost	This is the average cost per event/workshop for a project.	Integer
14.	Sub-Award or Grants	sub_grants	Sub-grants are funding mechanism to beneficiaries on a project such as start-ups, innovators or other beneficiaries being supported on a project. Whereas sub-ward is the primary grant recipient (the "prime recipient") engaging another organization (the "sub-recipient") to help implement the objectives of the original grant.	Integer
15.	Number of Events	num_events	This is the number of events/workshops plan to be executed on a project. High, medium, and low complexity project requires high, moderate, and relatively low people travelling respectively.	Integer
16.	Other Direct Project Expenses (ODC)	other_direct_expenses	There are expenses that are directly attributable to a specific project but do not fall under labor costs, direct materials other defined category.	Integer
17.	Total Project Cost	total_cost	This is the total cost of a consulting project including personnel cost, travels and other direct expenses.	Integer

2.2. Domain Knowledge Utilization

Domain expertise knowledge was involved in identifying and prioritising the factors influencing consulting projects. The process involved collaboration with subject-matter experts to ensure comprehensiveness of selected factors, categorisation of the identified factors into project-specific, client-specific, and industry-relevant financial data. The identified factors were validated by cross-reference with existing frameworks and methodologies in consulting project literatures or related studies.

2.3. Simulation Techniques

Due to lack of adequate dataset on consulting projects and data privacy, a simulation of the 80 real-world consulting projects was created for a controlled and structured data science approach for training machine learning models.

- **Simulation Design:** A simulation framework was developed utilizing a combination of rule-based, standard practices, industry contexts and programming to replicate the dynamic of real-world scenarios. This framework incorporated the identified factors, modelling their interactions and on project outcomes.
- **Data Collection & Generation:** Using the simulation frameworks, 2000 synthetic project scenarios were generated with the features listed in Table 2 above.
- **Assumptions:** Based on the insights from the investigation phase, the simulation process assumed a combination of linear and complex non-linear relationship between the 17 inputs: (16 independent variables) and 1 target variable (the project cost). The assumptions and insights from the investigation supported the identification and selection of 5 machine learning models from which the most suitable one for the regression task was selected.

2.4. Machine Learning Model Development

The generated dataset was utilized to train, validate, and test the performance and generalisation effectiveness of the machine learning models in cost estimation for consulting projects.

- **Feature Selection:** The features utilized in developing the ML models were derived directly from the identified influencing factors. All 17 features importance were analysed using Random Forest Regressor, Mutual Information and SHAP feature selection methods.
- **Model Training:** A supervised machine learning approach was adopted, with algorithms selected based on their suitability for the predicting the cost consulting project ('the problem')
- **Model validation and Performance/Success Metrics:** The goal is to minimize prediction error as cost prediction error can have significant financial and operational implications: both overestimation and underestimation can affect decision-making and profitability. Hence, the models were evaluated using multiple performance metrics for comprehensive assessment. These metrics include **Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Squared Error (RMSR) and R2 Score to assess predictive performance**.
- **Model Optimization:** After cross validation were carried out using the Random Forest Regressor to examine the models' performance and generalisation capability**;** the Grid and Random Search hyperparameters tuning process were both implemented to establish the best hyperparameters to deployed on the final test dataset for model's optimisation; ensuring that the models are neither under fitted or overfitted. Hence, improving predictive accuracy and computational efficiency.

2.5. Ethical Considerations and Limitations

The simulation techniques deployed in simulating real-world consulting project scenarios integrated ethical considerations in ensuring that no proprietary or sensitive from the investigated projects were disclosed. In addition, the limitations attributed to the simulation techniques were acknowledged, including reliance on the assumptions on the interactions and relationships amongst the features and oversimplification of real-world dynamics.

The research addresses the key stages involved in the implementation including **data acquisition, feature selection, model development, training, testing, and performance evaluation** as detailed in **fig. 1** below. This is in accordance with CRISP-DM model [12].

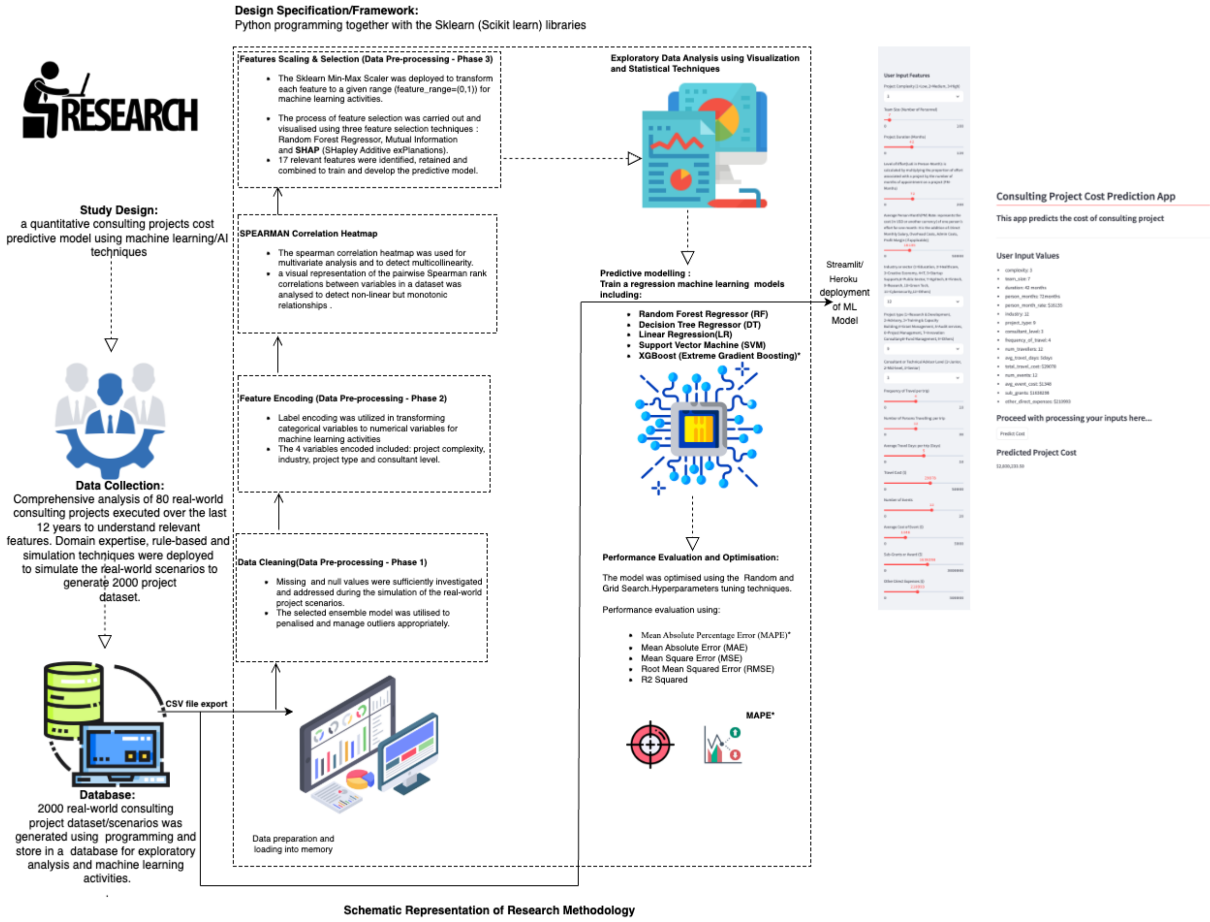

Fig. 1: Schematic representation of research methodology

3. Results

The primary objective of this research was to develop an AI/ML model to improve the prediction of consulting projects cost based on relevant factors identified from the investigation of 80 real-world consulting projects. The overall success metric of the ML predictive system is to reduce prediction errors to achieve reliable and accurate cost estimates. Hence, minimizing performance metrics such as Mean Absolute Percentage Error (MAPE) expresses error as a percentage and ensure the model meets practical needs for accuracy and usability. Other success metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R2 enhance the practical interpretability of the performance of models in terms of measuring average prediction error, penalizing larger and explanation of how well the models explains variance in the data respectively as shown in Table 3a and 3b. The dataset was divided into training (n=1400), validation (n=300) and testing (n=300) for building, training, validating, and testing the model respectively. The five ML/AL models in Table 3a were selected as appropriate model for the regression problem offer a mix of simplicity, interpretability, and predictive power for both linear and complex non-linear interactions in the variables. The generalisation capability of the models was evaluated using cross validation and best hyperparameters were selected using the Grid Search and Random Search hyperparameters tuning techniques as shown in Table 6a, 6b and 6c respectively.

The Ordinary Least Square (OLS) linear regression and mutual information hypothesis testing were used to test the hypothesis that the different factors have effect in predicting consulting projects cost ('target variable') as shown in Table 7 & Table 8 respectively. The Spearman correlation heatmap was deployed to investigate the multivariate non-linear, complex relationship and detect multi-collinearity amongst the 17 features in the dataset as detailed in Fig.3. The Random Forest Regressor, and Mutual Information and SHAP (SHapley Additive exPlanations) were used to gain insights into feature importance and selection of relevant features utilized to develop the ML models. in Fig.2, Table 4, Fig.6a and Fig.6b respectively.

Table 3a: Comparative Analysis of machine learning model results: using the default model without hyperparameters tuning

S/N	MODELS	MAPE	MAE	MSE	RMSE	R2 Score
1	Linear Regression	0.3495	4.324e+05	4.612e+11	6.791e+05	0.9127
2	Random Forest Regressor	0.3642	4.109e+05	4.228e+11	6.503e+05	0.9133
3	Decision Tree Regressor	0.406	4.071e+05	3.702e+11	6.084e+05	0.9223
4	XGBoost(Extreme Gradient Boosting)	0.09219	1.902e+05	9.38e+10	3.063e+05	0.9767
5	Support Vector Machine (SVM)	1.08	1.569e+06	6.008e+12	2.451e+06	-0.1595

Table 3b: Comparative Analysis of machine learning model results on final test dataset: using Best Estimator Hyperparameters from Random Search Hyperparameters Tuning

S/N	MODELS	MAPE	MAE	MSE	RMSE	R2 Score
1	Linear Regression	0.4226	4.417e+05	4.76e+11	6.899e+05	0.9063
2	Random Forest Regressor	0.3153	4.466e+05	5.119e+11	7.155e+05	0.8759
3	Decision Tree Regressor	0.3841	6.673e+05	1.443e+12	1.201e+06	0.7351
4	XGBoost(Extreme Gradient Boosting)	0.05077	1.041e+05	3.09e+10	1.758e+05	0.9937
5	Support Vector Machine (SVM)	0.9982	1.463e+06	5.132e+12	2.265e+06	-0.1804

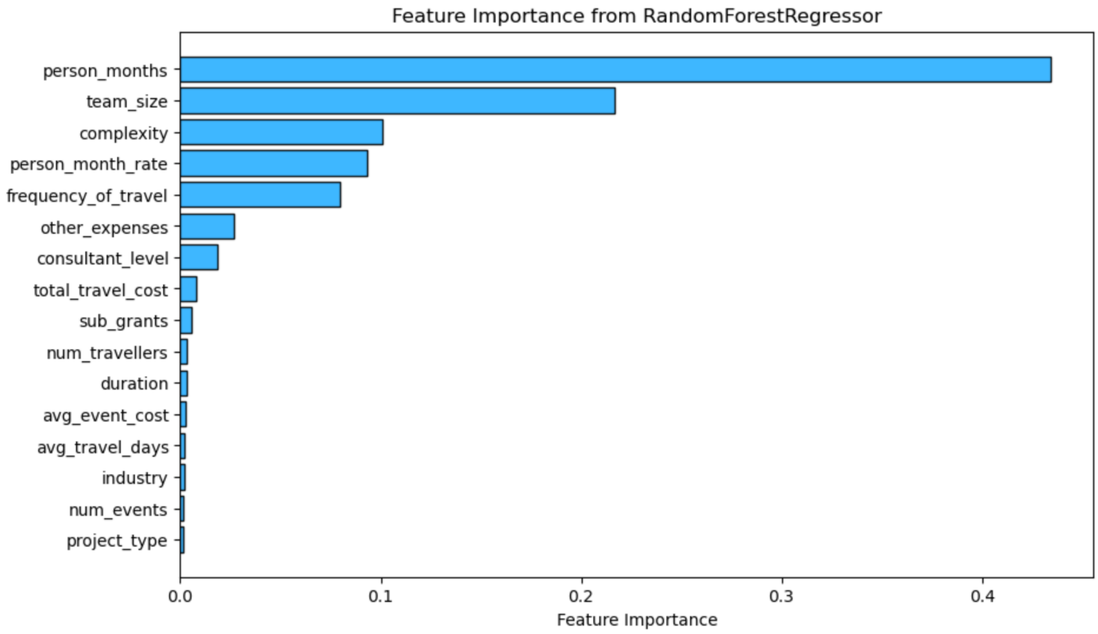

Fig.2: Feature Importance from RandomForestRegressor

Table 4: Feature Importance from RandomForestRegressor

S/N	Feature	Importance
1.	person_months	0.434028
2.	team_size	0.216964
3.	complexity	0.100870
4.	person_month_rate	0.093360
5.	frequency_of_travel	0.079924
6.	other_expenses	0.026675
7.	consultant_level	0.018252
8.	total_travel_cost	0.007712
9.	sub_grants	0.005582

10.	num_travellers	0.003320
11.	duration	0.003289
12.	avg_event_cost	0.002564
13.	avg_travel_days	0.002179
14.	industry	0.001805
15.	num_events	0.001751
16.	project_type	0.001726

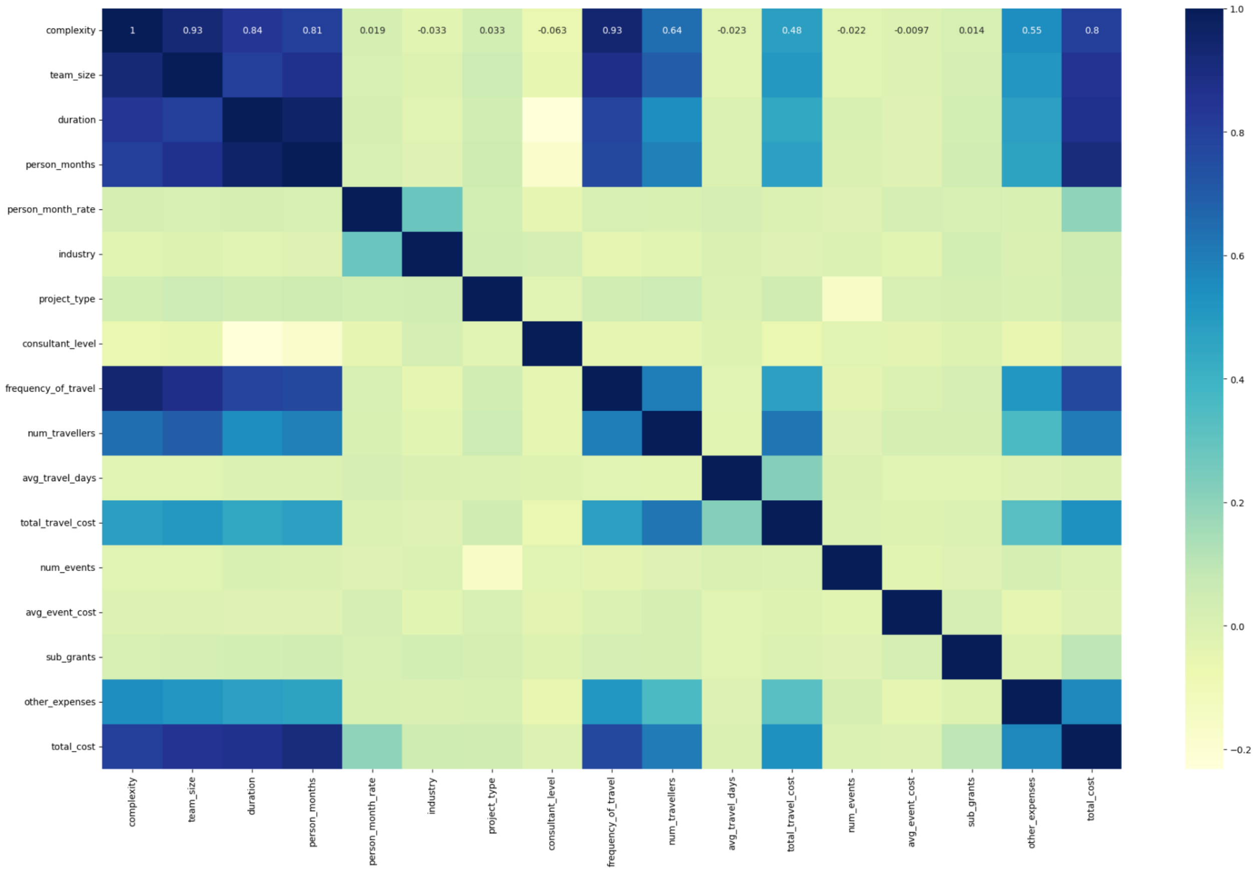

Fig3: Correlation Heatmap Showing Multivariate Relationships/Interaction Amongst Variables

Table 6

Table 6a: Cross Validation Results Using RandomForestRegressor: Kfold_split = KFold(n_splits=5)

Cross-validation scores: [0.95926443 0.96835152 0.95134855 0.95643566 0.95656239]
Cross val mean: 0.958 (std:0.006)
{'max_depth': 12, 'max_features': 12} Best Score: %s 0.959

Table 6b: Hyperparameters Tuning Using Grid Search: Fitting 5 folds for each of 2916 candidates, totalling 14580 fits

Hyparameters	Declared Dictionary of Hyperparameters for Tuning	Best Estimator Hyperparameters
n_estimators	[100, 200, 300]	'n_estimators': 300
learning_rate	[0.01, 0.05, 0.1, 0.2]	learning_rate': 0.1
max_depth	[3, 6, 10]	'max_depth': 3
min_child_weight	[1, 5, 10]	'min_child_weight': 5
subsample	[0.6, 0.8, 1.0]	‘subsample': 0.6
colsample_bytree	[0.6, 0.8, 1.0]	'colsample_bytree': 0.8
gamma	[0, 0.1, 0.2]	'gamma': 0

RMSE: 261608.9578486284
R-squared: 0.9873732781263932

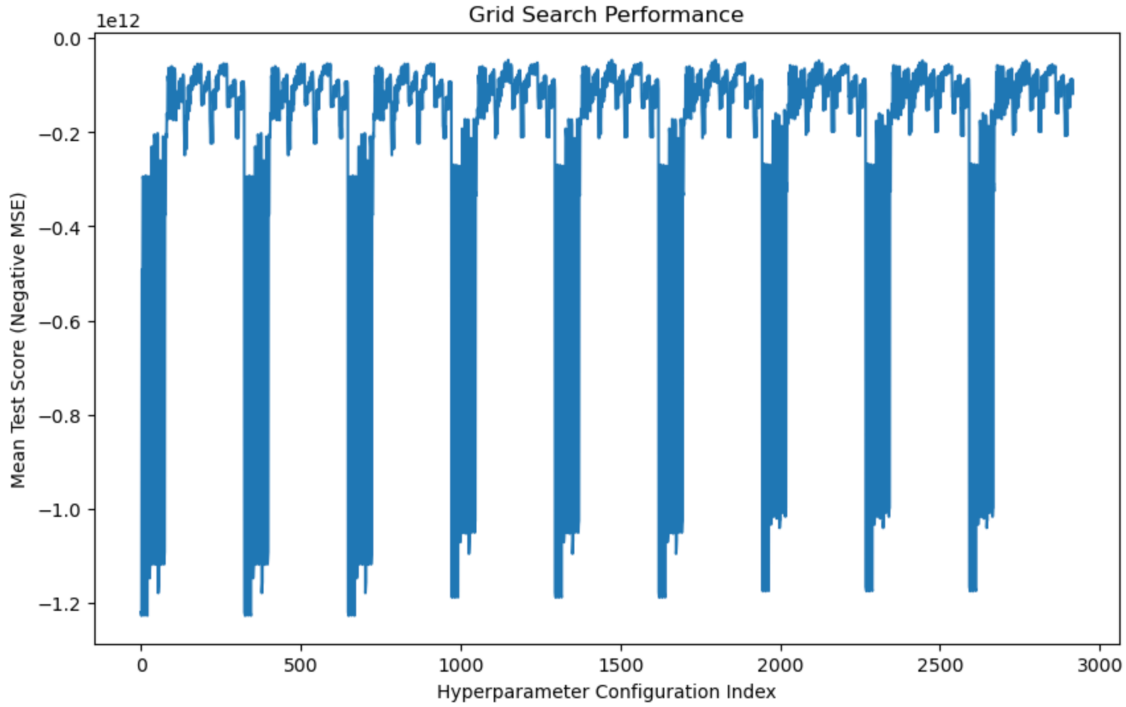

Fig4: Hyperparameter Tuning using Grid Search

Table 6c: Hyperparameters Tuning Using Random Search: Fitting 5 folds for each of 100 candidates, totalling 500 fits

Hyparameters	Declared Dictionary of Hyperparameters for Tuning	Best Estimator Hyperparameters
n_estimators	[100, 200, 300, 400, 500]	n_estimators: 500
learning_rate	[0.01, 0.05, 0.1, 0.15],	learning_rate: 0.1
max_depth	[3, 6, 10, 12]	max_depth: 3
min_child_weight	[1, 2, 5]	min_child_weight: 1
subsample	[0.6, 0.8, 1.0]	subsample: 0.8
colsample_bytree	[0.6, 0.8, 1.0]	colsample_bytree: 0.8
gamma	[0, 0.1, 0.2]	gamma: 0.2

reg_alpha	[0, 0.1, 0.5]	reg_alpha: 0.5
reg_lambda	[1, 1.5, 2]	reg_lambda: 2

RMSE: 188568.46924572595
R-squared: 0.9923151198860882

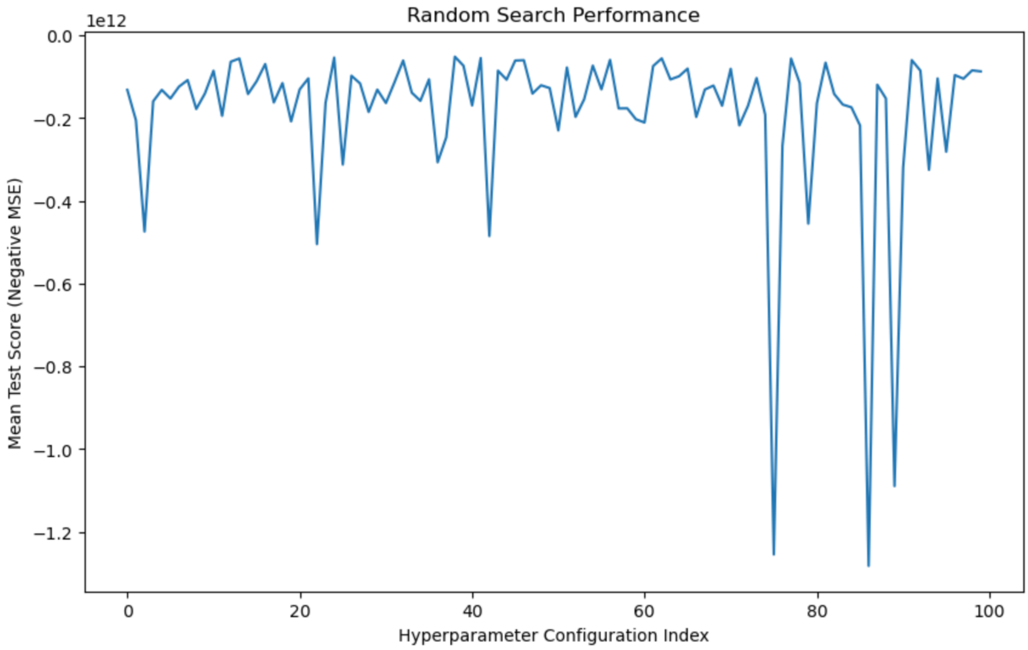

Fig5: Hyperparameter Tuning using Random Search

Table 7: OLS Regression Results Testing the Hypothesis of Significance of Interactions Amongst the Features

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             total_cost   R-squared:                       0.923
Model:                            OLS   Adj. R-squared:                  0.922
Method:                 Least Squares   F-statistic:                     1035.
Date:                Sun, 05 Jan 2025   Prob (F-statistic):               0.00
Time:                        14:40:33   Log-Likelihood:                -20604.
No. Observations:                1400   AIC:                         4.124e+04
Df Residuals:                    1383   BIC:                         4.133e+04
Df Model:                          16
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const               -2.285e+06   1.36e+05    -16.795      0.000   -2.55e+06   -2.02e+06
complexity            5.14e+04   1.06e+05      0.485      0.628   -1.57e+05    2.59e+05
team_size             2.01e+04   9255.143      2.172      0.030    1942.443    3.83e+04
duration            -2843.9022   5979.362     -0.476      0.634   -1.46e+04    8885.698
person_months        3563.0404    240.404     14.821      0.000    3091.445    4034.636
person_month_rate     317.2738     14.169     22.392      0.000     289.478     345.070
industry             5127.2007   4896.322      1.047      0.295   -4477.820    1.47e+04
project_type         -316.7061   6190.610     -0.051      0.959   -1.25e+04    1.18e+04
consultant_level     3.859e+05   2.16e+04     17.863      0.000    3.43e+05    4.28e+05
frequency_of_travel -2.575e+04    1.5e+04     -1.718      0.086   -5.52e+04    3652.286
num_travellers       9887.2376   4181.045      2.365      0.018    1685.361    1.81e+04
avg_travel_days      1157.1141   7262.192      0.159      0.873   -1.31e+04    1.54e+04
total_travel_cost       0.9535      0.080     11.883      0.000       0.796       1.111
num_events            329.8122   4346.300      0.076      0.940   -8196.241    8855.865
avg_event_cost         18.6522     18.780      0.993      0.321     -18.189      55.493
sub_grants              1.0778      0.111      9.700      0.000       0.860       1.296
other_expenses          0.9573      0.052     18.237      0.000       0.854       1.060
==============================================================================
Omnibus:                      394.736   Durbin-Watson:                   1.959
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3945.683
Skew:                           1.014   Prob(JB):                         0.00
Kurtosis:                      10.971   Cond. No.                     6.11e+06
==============================================================================
```

Table 8: Mutual Information Hypothesis Testing and Features Selection: To Establish Non-Linear/Complex Interactions Amongst the Features

S/N	Features	Mutual Information
1.	person_months	0.897651
2.	team_size	0.888145
3.	frequency_of_travel	0.872811
4.	complexity	0.872112
5.	duration	0.829419
6.	other_expenses	0.465827
7.	total_travel_cost	0.449121
8.	num_travellers	0.341307
9.	sub_grants	0.088576
10.	person_month_rate	0.033850
11.	avg_travel_days	0.017327
12.	industry	0.013781
13.	num_events	0.006355
14.	project_type	0.000000
15.	consultant_level	0.000000
16.	avg_event_cost	0.000000

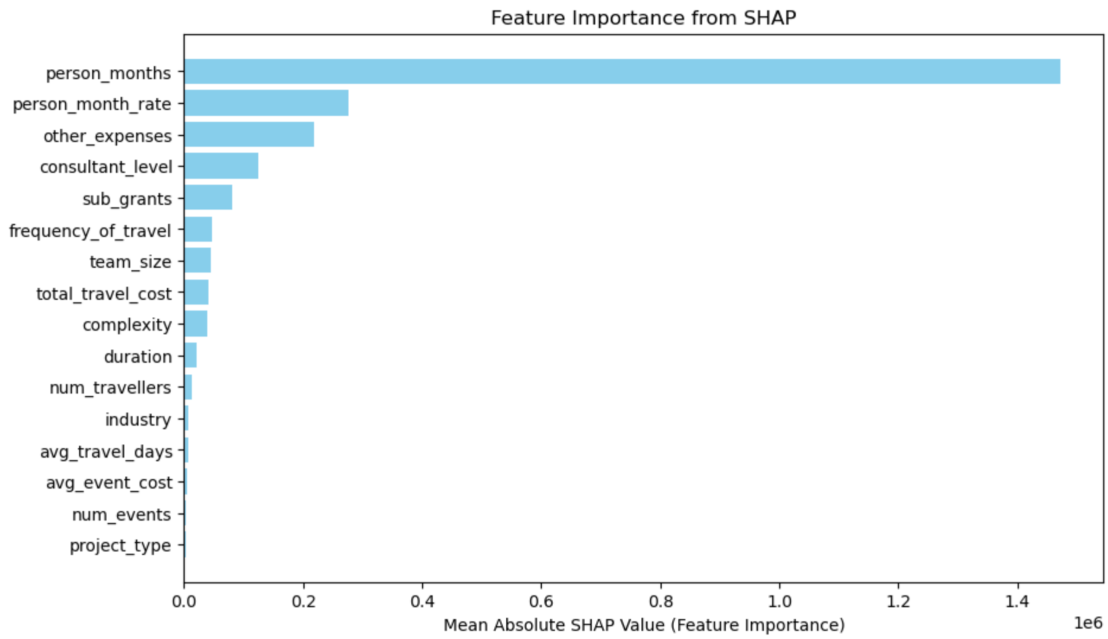

Fig. 6a: Features Importance from SHAP

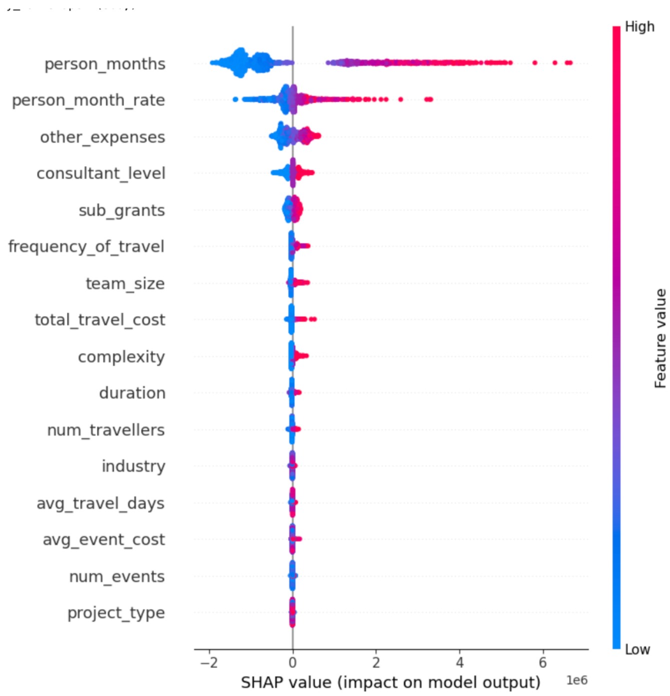

Fig. 6b: Features Importance from SHAP

4. Discussions

In this study, five (5) machine learning models were deployed and evaluated in predicting the cost of consulting projects, a significant decision for firms aiming to optimize their bidding strategies and resource allocation. The ML models deployed include Linear Regression, Random Forest Regressor, Decision Tree Regressor, XGBoost (Extreme Gradient Boosting) and Support Vector Machine (SVM). The models were trained, validated, and tested on 2000 dataset generated from simulation of real-world consulting projects in different domains such as education, healthcare, creative economy, information technology (I.T), start-up supports, public sector, agrictech, fintech, research consultancy, green technology, cybersecurity amongst others. The result indicates that XGBoost model is the most suitable machine learning model with the least prediction error across diverse consulting domains: MAPE, MAE, MSE, RMSE and R2 Score of 0.05077(5%), 1.041e+05, 3.09e+10, 1.758e+05, 0.9937(99%) respectively. The MAPE of 5% validates that XGBoost outperforms the traditional baseline linear regression with MAPE of 42% by a significant reduction of prediction error of approximately 37%. Notably, the XGBoost model perform well on consulting projects with high complexities that require more human, material resources and where historical variability costs are high.

The Mutual Information (MI) was utilized to rank each features by their relevance to the target variable ('project cost'), the top ten relevant features that contributed to the performance of the XG-Boost Model are: **person-month, team size, frequency of travel, project complexity, duration, other direct expenses, total travel cost, number of travellers, sub-grants/awards, and person-month (PM) rate**. The significantly high MI value for person-month, team size, frequency of travel, project complexity and duration of (0.897, 0.888, 0.873, 0.872 and 0.830) respectively implies that the combination of these 5 independent variables contain significant information about the target variable ('the project cost'). Person-months holds the most information with an MI value of 0.897: a stronger driver of project cost. These results are also validated the Random Forest Regressor and SHAP model features importance in Fig. 2, 6a and 6b respectively. Furthermore, team size, frequency of travel, project complexity and duration complements provided these features integrate other influencing elements of consulting project cost. Hence, Person-Months when combined with the PM rate determined the base cost of any consulting project before accounting for other direct project expenses.

4.1. Performance Evaluation of the ML Model

In line with the goal of the study and ease of interpretability of the models for the various stakeholders of the research, the primary performance measurement (success metric) of the model was based on the Mean Absolute Percentage Error (MAPE): expresses error as a percentage and ensure the model meets practical needs for accuracy and usability. The MAPE of 5% for the XGBoost model indicates that, on the average, the model's prediction is off by only 5% relative to the actual project cost. Hence, this is generally considered a good level of accuracy with minimum predictive error when compared to the MAPE of the other 4 models: Linear Regression, Random Forest Regressor, Decision Tree Regressor, and Support Vector Machine (SVM) with MAPE of 42%, 32%, 38%, 99% respectively.

The XGBoost model performance was evaluated on actual real-world scenario data points against the predictions as summarized in Table 10 below. The comparative analysis shows that for most of the projects, the model predicts the cost within 5% higher or lower than the actual project cost.

Table 10: Comparative Analysis of the Predictive Performance of the XG-Boost Model

Row Indices	Project Complexity	Actual: Train Dataset (Project Cost) in USD$	Predictions: Test Dataset (Project Cost in USD$)
771	1=Low	208,310.4	289,765.34
1106	2=Medium	217,3369.2	206,3641.2
787	2=Medium	1,165,910.4	1,108,702.9
1788	3=High	5,021,000.8	5,021,526.0
785	3=High	4,766,280.0	4,735,031.0

The above comparative of actual data and predictions demonstrated the model's accuracy and it effectiveness in minimizing prediction errors. These predictions enabled consulting firms to present a competitive and profitable proposals while managing clients' expectations effectively.

4.2. Validation of Hypothesis, Features Importance, and Interactions

The results of the ordinary least square (OLS) regression model in Table 7 were used to test the level of significance and relationship of each of the 16 independent variables and the target variable (project cost). The p-value was used to establish the threshold, and level of significance of each variable. 8 of the 16 independent features have p-values equal to zero but less than 0.05 (alpha: critical value). These features include person-months, person-month (PM) rate, team size, consultant level, number of travellers, total travel cost, sub-grant/awards, and other project expenses. They also have a positive relationship with the target variable and relatively low standard error: indicating that they are statistically significant. Although the remaining 8 features have p-value higher than the threshold of 0.05, but they have a significantly high co-efficient and some have a positive relationship with the target variables making them to have explanatory power or holding relevant information about the total project cost. Therefore, we can reject the null hypothesis (H_0) and accept the alternate hypothesis (H_1) that the predictors (independent variables) are statistically significant. In addition, the linear regression was used to evaluate the magnitude of variance in the target variable explained by the predictors; this produced R^2 and adjusted R^2 of 92.3% and 92.2 respectively, making the interactions amongst the features statistically significant.

4.3. Model Optimisation

In the study, cross validation was combined with rigorous hyperparameters tuning using Random and Grid Search. This process of fine tuning the hyperparameters was necessary during the training, validation and testing of the ML models [8]. Cross validation was used to evaluate the perform of the model and how it generalise on unseen test dataset. Whilst the grid and random search techniques provided a systematic and robust approach to identifying the optimal configurations that generalises well to new datasets. The mean cross validation best result using 5 K-fold splits was 95%, Table 6a. The best accuracy achieved by grid search with high iterations and computational resources was 98% with RMSE of 261,609, Table6b. Whereas random search achieved an accuracy of 99% with RSME:188,569 with less computational resources, Table 6c. Hence, the random search was deployed on the XGBoost model because it was well suited: offers greater scalability, efficiency for high-dimensional and complex task such as predicting the cost of consulting projects.

4.4. Main Findings

The comparative analysis of the actual versus predicted project cost of the XGBoost model results in Table 10, Mutual Information (MI), Random Forest Regressor, and SHAP feature importance in Table 4, Fig.2 , Table 8, and Fig.6a respectively; all indicated that person-months ('level of human effort') devoted to a project emerged as the strongest explanatory feature with (MI score of 89.7%) and holds significant information about the target variable ('total project cost'). This MI score was higher than other features such as team size (MI:88.9%), frequency of travel (MI:87%), project complexity (MI:87%), project duration (82.9%). Furthermore, the findings of the research show that:

- Project with **HIGH** complexity requires a much higher person-month, team size, frequency of travels, relatively high duration, and other expenses.
- Project with **MEDIUM** complexity requires a relatively moderate person-month, team size, frequency of travels, duration, and other expenses.
- Project with **LOW** Complexity requires a relatively low person-month, team size, frequency of travels, duration, and other expenses.

Previous studies in consulting projects cost prediction have heavily relied on traditional historical cost estimation such a Linear Regression which is limiting in capturing non-linear complexities associated with most consulting projects. In addition, they have overlooked project complexities and person-months leading to under/over estimation of project cost and non- competitive bidding strategies. The findings and results of this research indicate that organisations can leverage project complexity metrics for better project planning, resource allocation, competitive yet profitable bidding strategy, risk management, operational efficiency, and ultimately accurate cost estimation. The insights from the research can be integrated into consulting firms project management system for data-driven cost estimate automation.

Additionally, this research incorporates larger and more diverse project dataset (n=2000) and addressed the challenges of lack of quality and availability of dataset for ML/AI research in predicting consulting projects cost, offering more generalisable insights compared to previous studies. While this study contributes and enhance existing studies on the applications of ML/AI in predicting the cost of consulting projects, it shares limitations with existing studies, the dataset did not include real-time changes in project which could have implications on cost predictions. Hence, future work could integrate dynamic cost prediction models that reflect time-series project dataset.

5. Conclusion and Recommendations

This research clearly demonstrated the explanatory power of different features when combined in predicting the cost of consulting projects. Person-month emerged as strongest predictive feature and hold significant information about the target variable (‘project cost) with MI of 89%. When combined with other relevant factors such as project complexity, team size, duration of projects, frequency of travels amongst others, it can lead to improve project cost prediction and resource optimisation. The Gradient Boosting model outperformed traditional linear regression techniques with an MAPE of 5% demonstrating the effectiveness of non- linear modelling in capturing the complexities commonly associated with consulting project features and costs. The research leverage advanced machine learning models with identified relevant features such as person-months, project complexity, team size, project duration and frequency of travel as the critical cost drivers influencing consulting project costs. Hence, addressing the limitations associated with traditional cost estimation and provides actionable insights to service providers, clients, project managers and other beneficiaries of the research.

Despite the decent performance of the model utilized in this research, it utilized a static historical project dataset which does not integrate changes in project scope, client satisfaction, economic and market data to enhance the model performance and generalisation; highlighting the need for future work to explore real-time and more dynamic cost predictive model or decision support system.

Consulting firms and other key stakeholders of a project should integrate a standardized framework during the project initiation and planning stage to fully understand project complexity, do a proper categorisation into LOW, MEDIUM, or HIGH to identify critical features such as person-month, team size, duration and uncovers other relevant features, drivers, or requirements for improved cost estimation practices.

Summary Table

What was already known about the topic:

- Machine learning/AI in predicting the cost of consulting projects is an emerging area of research and practice.
- The current state of art and previous studies have heavily relied on traditional cost estimation such as historical benchmarking and expert judgement, overlooking the impact person-month (‘human effort’), project complexity, duration, and other relevant factors.
- A lack of standardized datasets and metrics for consulting projects makes it difficult to compare findings across projects.

What this study added to our knowledge:

- Beyond using historical benchmarking and expert knowledge, the study has proven that person-month (‘human effort’) is the strongest feature to accurately predict the cost of consulting project. When combined with other factors such as project complexity, duration, team size, it enhances the accuracy of cost prediction.
- The study also proposed standardization of cost estimation metrics. In particular, a structured data science approach to quantifying project complexity into LOW, MEDIUM, or HIGH based on level of effort (person-months) and other resources requirements.
- The study also deployed an explainable and transparent AI Model such as XG Boost not only to predict cost, but also provides a practical explanation of the predictions.
- Provided practical and novel decision support system to consultants, clients, and other stakeholders to accurately predict consulting projects and integration of the insights to their bidding strategies.
- The project also combined domain expertise, rule-based and simulation of real-world project scenarios to generate 2000 dataset. The has contributed to the creation, and sharing anonymised, large-scale consulting project dataset for future work, performance improvement, training, and validation of ML models.

Acknowledgements
I am heartily thankful to Co-Creation Hub Limited and Gauge Innovation Limited for the opportunity to examine real-world consulting project scenarios that have made this work successful.

Data Availability Statement
The project dataset analyzed during the investigation phase of this study was generated from real-world scenarios. Due to the proprietary nature of the real-world data and the confidentiality agreements with participating organizations, the dataset is not publicly available. However, simulation code and rule-based logic files used in generating the 2000 synthetic dataset utilized for training and developing the ML models are available upon reasonable request to the corresponding author.

Ethical Statement
Domain experts who contributed to this study were fully informed of the research goals and provided explicit consent for their knowledge to be used in this work and were given the opportunity to withdraw from the research at any time, without reason. In addition, the study was conducted in accordance with the ethical standards outlines by the participating organisations. Efforts were made to minimize bias in both real-world data collection and simulation processes to ensure fair and equitable outcomes. The study did not engage any less privileged, physically challenged individuals nor animal subjects.

Ethics Approval
The actual dataset utilized in this research to train, validate, test, and develop the machine learning models were generated using rule-based and simulation techniques to model real-world project scenarios. This contains no personally identifiable information and completely anonymized. However, Ethical approval was applied and granted for this research by Gauge Innovation Limited.

Conflict of Interest Statement
The author declares no conflicts of interest related to this research. This work was conducted independently and was not influenced by any commercial, financial, or organizational relationships that could be construed as a potential conflict of interest.

6. References

[1] M. Nawaz, M. Kumar, M. Zaki, S. Gouda, P. Raj, N. Kumar, and R. Das, A Methodology for Software Cost Estimation Using Machine Learning Techniques: International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020) (July 10, 2020). Available at SSRN: https://ssrn.com/abstract=3647639 or http://dx.doi.org/10.2139/ssrn.3647639 (Accessed: 13 October 2024).

[2] H. Ming-Wei, D. Nicholas, and P.K Senyo, Applied Algorithmic Machine Learning for Intelligent Project Prediction: Towards an AI Framework of Project Success (2021). Advanced Project Management, 21(4), Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3823900 (Accessed: 15 October 2024).

[3] X. Gao, M. Khorasani, S. V. Krishnan, L. Nittala, and S. Wang, A Machine Learning Framework for Project Performance Prediction, Interpretation, and Inference (December 09, 2024). Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5051652 (Accessed: 20 December 2024).

[4] R. Dzhusupova, V. Shteriyanov, J. Bosch, Jan and H.H. Olsson, Revolutionizing Engineering and Construction Projects: The Role of Artificial Intelligence in Cost Estimation and Procurement. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4873615
(Accessed: 25 October 2024).

[5] S. Farzaneh, M. Zadeh, M. Sadat, H. Ghafourian and E. Zadeh, Applications of Machine Learning in Financial Accounting for Industrial Engineering: A Case Study on Cost Estimation and Forecasting (October 15, 2024). Available at SSRN: https://ssrn.com/abstract=4991489 or http://dx.doi.org/10.2139/ssrn.4991489 (Accessed: 30 October 2024).

[6] C. Kan, Z. Yuezheng, H, Naixin, Y. Chao, M. Ji and Z. Tong, Cost Prediction for Water Reuse Equipment Using Interpretable Machine Learning Models. Available at SSRN: https://ssrn.com/abstract=4712568 or http://dx.doi.org/10.2139/ssrn.4712568
(Accessed: 1 November 2024)

[7] G. Hopper, Deep finance: corporate finance in the information age. First ed., United States: Leaders Press, 2021. pp.25-51.

[8] A.C. Muller, S. Guido, Introduction to Machine Learning with Python. First ed., Sebastopol: O'Reilly Media, Inc, 2016.

[9] Scribbr. Research Methods: Definitions, Types and Examples. https://www.scribbr.co.uk/category/research-methods/, 2022 (Accessed 15 November 2024).

[10] European Commission, What is 'Human Effort' (person-months) and how to Calculate it? Available at: https://ec.europa.eu/newsroom/just/items/643967#:~:text=To%20calculate%20the%20total%20person,of%20months%20of%20the%20appointment (Accessed: 1 March 2024)

[11] Sploro, How to Calculate the Person Month (PM) Rate for your Horizon Proposal Budget. Available at: https://sploro.eu/how-to-calculate-the-pm-rate-for-your-horizon-proposal-budget/ (Accessed: 15 June 2024).

[12] Data Science Process Alliance, What is CRISP DM? Available from: https://www.datascience-pm.com/crisp-dm-2/, 2022 (Accessed 10 December 2024).