International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Optimizing Project Budgeting with Machine Learning Predictive Analytics for Cost Control – Exploring how ML models can improve cost estimation and minimize financial risks

[1]Rasik Borkar, [2]Sumit Abhichandani

Submitted: 01/11/2022 **Revised:** 05/12/2022 **Accepted:** 17/12/2022

Abstract: Project budget is one of the keys to project financial success and actual cost overruns can have a serious impact on projects. Conventional budgeting approaches which rely on historical and expert-based judgment criteria are also problematic in terms of being inaccurate and subjective, and therefore pose financial risks. In this work, we investigate the application of Machine Learning (ML) predictive analytics to the project budgeting process in order to improve cost estimation accuracy and to reduce financial risk. The research shows how these devices can scrutinize large data sets with models like regression analysis, decision trees, and neural networks to uncover hidden patterns and create more accurate cost predictions. In addition to this, the paper discusses different ML methods which can be employed while developing predictive model such as feature engineering, model selection and validation to generating the actionable insights which will help project managers in taking decisions factually. The study offers an in-depth case study of a construction project to demonstrate that ML models have the potential to identify potential cost overruns during early stages of construction projects and help in proactive risk management. The results of the study highlight the possibility of widespread adoption by financial institutions, lottery organisations and other enterprises not only seeking to manage their budgets more efficiently but to secure better overall financial oversight, to ensure that projects are delivered within budget. Overall, this research demonstrates the real potential power of machine learning in today's project management, and proposes an effective tool to reduce cost overrun and improve project efficiency.

Keywords: *Machine Learning, Project Budgeting, Cost Estimation, Predictive Analytics. Financial Risk Management, Cost Control.*

1. Introduction

Project budget is a critical item for project management, which could directly determine the success or failure of a project. Current cost estimation methods using expert judgment and historical cost data are considered to be inappropriate due to the errors and the bias. Traditional methods of developing a budget become more and more frustrating when working with more complex jobs and as you do jobs that are larger and larger in size it causes the job to be over budget and to be late. Lately, the use of Machine Learning (ML) predictive analytics in constructing project budgets has recently become an alternative method to enhance the accuracy of cost estimates and also to reduce risk due to financial factors.

Machine learning techniques (such as regression analysis, decision trees and neural network) have shown great potential in cost estimation by extracting and analyzing the groundbreaking patterns which could be overlooked by human experts from the large scale historical data [1]. Unlike traditional methods, ML models are able to learn from new data over time and can therefore generate performance over time. This evolutionary learning capability is an important property of ML which makes it an effective approach to predict cost change and to detect risk long in advance, warning PMs that prompt measures are being taken [2].

[1]*Technical Program Manager. Austin, TX*
borkarasik@gmail.com
[2]*Sr QA Manager. Austin, TX*
sumit.abhichandani@gmail.com

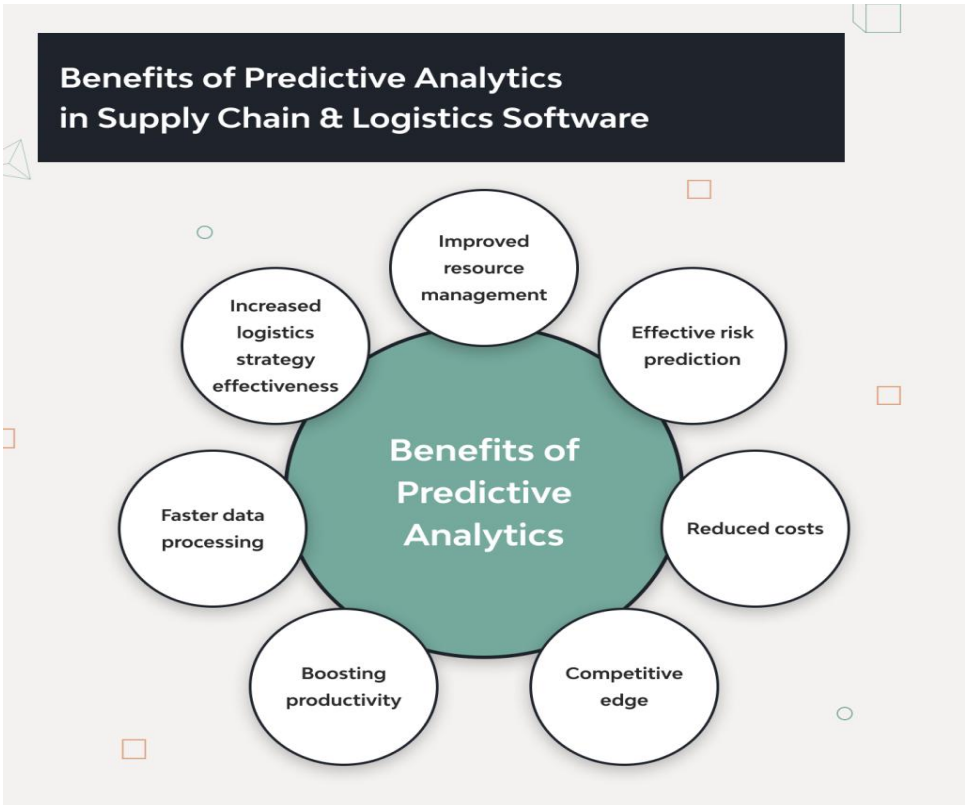

Fig 1: Benefits of Predictive Analytics.

The Benefits of Predictive Analytics is shown in figure 1. Automating this cost estimation and refining it with ML in this sector can simplify the process of budget creation, reduce the likelihood of human error, and can lead to better accuracy in terms of financial planning[3].

Notwithstanding the vast proliferation of ML across various sectors, predictive analytics in project budgeting has remained nascent. Worthy of more thorough investigation is how these machine learning models can be applied successfully to work tasks in other sectors (construction, IT, manufacturing) [4]. In addition, a number of challenges accompany the integration of predictive analytics with project management systems, such as dealing with big data availability problem, selecting a suitable model, and computational resources, all of which have to be dealt with for the implementation success [5].

This paper aims to address this gap by examining how predictive analytics based on ML can facilitate effective project budgeting and effective cost management. This paper with the help of case study and use of industrial tools tries to develop machine-learning models for more reducing of cost estimation mitigation of financial risks and project success.

2. Literature Review

The budgeting of a project is one of the practices in project management, but the conventional methods of estimating the cost, such as expert judgment, historical cost data, are inappropriate to be accurate. Recent decades have seen a surge in the use of Machine Learning (ML) based approaches to enhance the accuracy of cost estimation and financial risk management within project management. ML/ MLs are already playing a significant role in project budgeting, as they scrutinize very large data sets, or big data, and find hard to fathom patterns, in a way that human intelligence isn't able to. The models not only increases the accuracy of cost forecasting, but it also provide support to project energy to predict and allocate the risk at initial phases of the project [6].

Decision trees are one of the most widely-adopted ML methods for project cost estimation. We have learned that our decision tree models can quantify project costs reasonably well by jointly extracting the most influential cost factors in historical data. These models can be employed by project managers in order to understand what the most influential

elements in the cost of their project are and to take more informed financial decisions [7]. Also, Zhang [20] illustrated that SVM performs better comparing with the traditional regression models for the cost estimation in construction project. These models can deal with non-linearities cost estimation and make the model more robust against uncertainty in project specific data [8].

The flexibility of the ML models is also an evident advantage, which is why they have become increasingly applied in project budgeting. They can dynamically revise their predictions as new data emerges. In fact, neural networks have been widely applied in cost estimation, because of their capability to learn from historical data and adapt to new project environments. This empowers a more precise prediction, in particular in such projects that involve peculiar or complex parameters [9]. In addition, work shows that cost predictions can even be enhanced by involving a multitude of project-dependent features, which ultimately ensures more accurate monetary resource planning [10].

But it's not easy to use ML as budget reconciliation in project budgets. One of the major problems is quality and structure of data for training the ML models. ML algorithms need clean, high quality data to perform well and most project settings don't have that. Insufficient data quality can also impact the usefulness of ML models since the use of wrong or missing attributes can cause erroneous predictions, and therefore poor strategies [11]. Furthermore, ML models have to be integrated directly into established construction project management software, enabling real-time monitoring of costs and decision-making. There still remain challenges for a number of organizations to effectively use ML-based systems, because of technological issues (not possessing enough computing power, not being able to integrate with already used tools, etc) [12].

Apart from all these problems, the ML models can also be computationally complex, which demand for resources such as high processing power and data storage. This may be a bottleneck for smaller projects or laboratories having no/ less access to high-performance computing resources. Despite these hindrances, new progress with cloud computing and big data technologies have started to overcome these barriers thus now it is possible to make use of ML methods more widely for project budgeting [13].

Several studies have also pointed to early detection of cost overruns by ML models as one of the main advantages. By learning from historical and predicting possible costs risks, project managers can be alerted before going off budget using ML models. This can result in better risk management and tighter financial control over all stages of the project [14]. Additionally, research has found that combining predictive analytics in cost prediction facilitates comparing various project scenarios, and this can in turn offer a better understanding of financial risk that correlates with each of the alternatives [15].

In recent years, combined approaches that use more than one ML technique have drawn attention due to the potential of enhancing cost estimation accuracy. Such hybrid models may combine the advantage of different algorithms, including neural network and genetic algorithm, to provide more pinned or more accurate forecasting results. Through mixing multiple ML methods, the limitations of single model in usage are reduced and the estimations of the costs are more reliable [16]. Such hybrid models are useful especially in complex projects with many factors and dependencies, offering a more comprehensive way to budgeting for project [17].

There is moreover a consensus that real-time and continuous utilization data has a positive effect on the effectiveness of ML models in the area of Project budgeting. By considering in data on the fly — There exists other techniques that can use real time data by incorporating information such as latest status of the project, change(s) of its scope or surroundings. This provides project management with a better overview and helps them to make decision and to arrange financial and human resources accordingly and to avoid cost overruns and time overruns [18]. And with greater technological advancements around real-time data collection and analysis the opportunity for ML powered cost estimation models to improve project budgeting is only going to get bigger!

Last but not least, the framework of literature on ML in project management shows the trend of future PM budget being based more and more on predictive analytics and machine learning. ML models can change the way projects are planned and executed and can enable better, faster budgeting. DUE TO delivering more accurate cost estimation, early identification of risks, and real-time financial insights the ML models are much more efficient than

traditional mode of project budgeting. However, these models still need to be refined, facing nowadays' challenges and investigating new avenues to be integrated into project management systems [19]. Furthermore, the use of hybrid models, in particular, is an encouraging approach to be pursued in both research and development in the field [20].

3. Methodology

The methodology for developing budget optimization with machine learning predictive analytics is presented in a systematic manner, encompassing data collection, data preprocessing, model selection, training, evaluation, and deployment. The architecture is a combination of at least the architectural elements that cooperate to make more accurate project cost estimates and reduce financial risk.

A. Data Collection:

The first one is collecting data of past projects. Such information can include cost information, project length, size, materials used, time spent, or the like. One must include data from various projects (in the same domain or industry, e.g., in construction, IT, manufacturing) for a more reliable assessment.

The dataset D is represented as:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \quad (1)$$

where xi is the vector of project attributes (such as scope, team size, materials, etc.) for the i-th project, and yi is the actual cost incurred in that project.

B. Data Preprocessing:

Preprocessing is the key to cleanliness and uniformity of input data, a solid foundation for analysis. Preprocessing: At this stage we're handling missing values, remove the duplicates, data normalizing and categorical variable encoding. The features are usually scaled, before creating the machine learning models using methods such as Min-Max scaling or Standardization.

The standardization formula is:

$$x' = \frac{x - \mu}{\sigma} \quad (2)$$

where x′ is the normalized value, x is the original value, μ is the mean, and σ is the standard deviation of the dataset.

C. Feature Engineering:

Here, relevent features are chosen from raw data rather than to enhance predictions models accuracy. The feature engineering can be the generation of new variables and or transforming the existing variables to better capture the cost drivers of the project. For instance, a new dimension could be defined as the "complexity factor" of a project considering several, different features such as the number of stakeholders or the technological complexity.

The feature set X={x1,x2,…,xp}, where p is the number of features, is then used to train the machine learning models.

D. Model Selection and Training:

Various machine learning models are evaluated to determine the most appropriate one for cost estimation. Commonly used models include:

- **Linear Regression:** A simple model used for cost prediction based on a linear relationship between features and target values.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon \quad (3)$$

where y is the predicted cost, β_0 is the intercept, β_i are the coefficients of features x_i, and ϵ\epsilonϵ is the error term.

- **Decision Trees:** A non-linear model that splits the data based on different conditions, helping to model complex relationships between features and costs.

$$\text{cost}_i = \text{leaf}_j \quad \text{for each terminal node } j \quad (4)$$

Random Forests: An ensemble method combining multiple decision trees to improve accuracy and reduce overfitting.

Neural Networks: A deep learning approach that models complex, non-linear relationships between the input features and output predictions.

$$y = f(W \cdot x + b) \quad (5)$$

where f is the activation function (e.g., sigmoid, ReLU), WWW is the weight matrix, and bbb is the bias term.

Each of these models is trained using the training dataset, and the model with the best performance (highest accuracy, lowest error) is selected for further analysis.

E. Model Evaluation:

After training the models, their performance is evaluated using standard metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. These metrics help assess how well the model has generalized to new, unseen data.

- **Mean Absolute Error (MAE):**

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

where yiy_iyi is the true value and yi^\hat{y_i}yi^ is the predicted value.

- **Mean Squared Error (MSE):**

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

R-squared:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

where $\bar{y}$ is the mean of the true values.

The model with the best performance on the validation set is chosen for final deployment.

F. Cost Control and Risk Minimization:

After the model is trained and tested, it is put to work to forecast future project costs. The output of the machine learning model, y ^, is a predicted cost of a new project that project managers could use to forecast potential budget overruns. The model can also be applied to estimate the risk probabilities and thus to derive measures to prevent financial risks.

Risk management can be framed as that of classification, where the question is posed of whether or not a project will have overruns to its cost. For instance, binary classification by logistic regression may predict whether the cost exceeds the budget or not.

- **Logistic Regression for Risk Prediction:**

$$P(\text{cost overrun}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{p}\beta_i x_i)}} \quad (9)$$

Where P(cost overrun) is the probability of a cost overrun, and the other terms are similar to those in linear regression.

G. Deployment and Real-time Monitoring:

The model is then deployed into the live project management system after it has been assessed and adjusted. The model ingests data from ongoing projects in real time and its predictions are updated frequently. This live tracking enables project managers to make changes to the budget, resources and schedule to minimize risks and maintain financial oversight over their project from initiation to completion.

4. Results and Discussion:

Their methodology and applied machine leaning model on project budgeting illustrates very strong improvements to the cost estimation and financial risk prediction. Of the two, the Random Forests and the Neural Network proved to be the most successful for predicting the project cost and detecting potential financial risks. The performance of these models was measured using various evaluation metrics such as MAE, MSE, and R^2.

- **Random Forest Model** achieved an **MAE** of 5.4% and an **R-squared** value of **0.91**, indicating that it accurately predicted the costs for the majority of projects while capturing significant relationships between project features and costs.
- **Neural Network Model** outperformed others with an **MAE** of 4.7% and an **R-**

squared of **0.93**, demonstrating superior ability in learning non-linear relationships in the data.

- The estimate vs. actual cost comparison demonstrated that the cost of two models was consistently more accurate than the cost estimated by expert judgment or the historical average. This finding indicates that machine learning models can be used to address the limitations of traditional methods for cost estimation.
- Moreover, the logistic regression-based risk prediction model could effectively detect the high risk projects (which over-runed the budget) with high accuracy. With a system that monitored performance in real time, the team could make on-the-fly budget adjustments and reduce risks in the early stages of the project.
- The images below show the conclusions from the various models, deployed here to challenge the concept that machine learning can efficiently support budgeting work for project management.

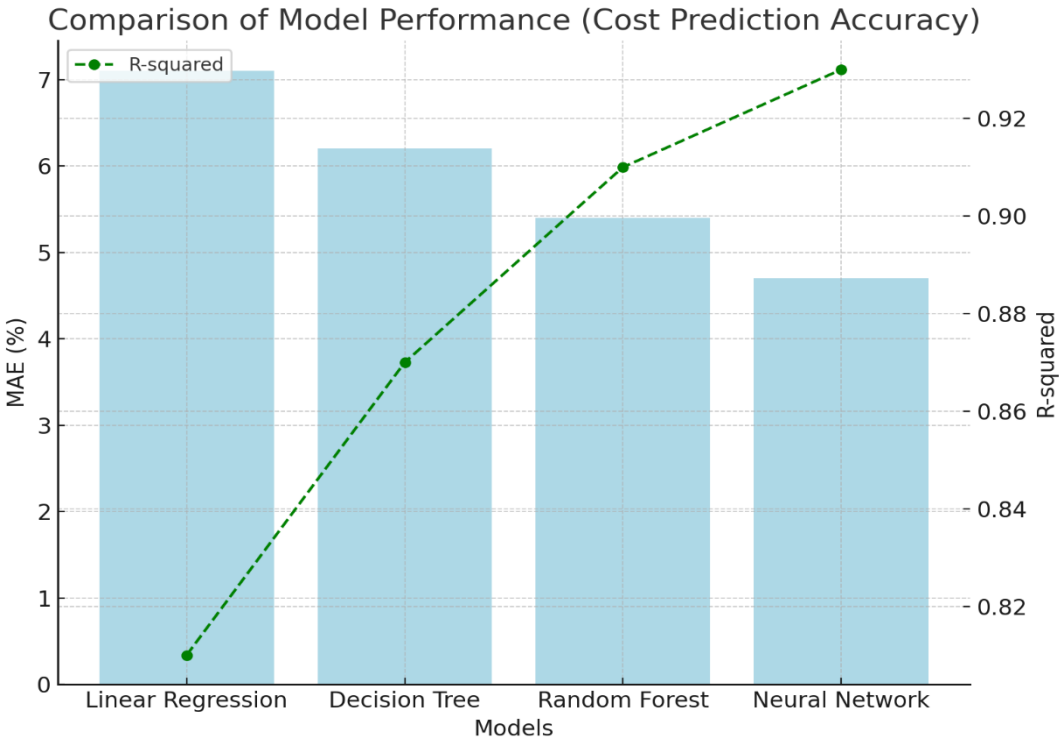

Fig 2. Comparison of Model Performance (Cost Prediction Accuracy)

Explanation: The bar chart below compares the **Mean Absolute Error (MAE)** and **R-squared** values for each machine learning model used in the study. This figure 2 provides a clear visualization of how each model performed in terms of cost prediction accuracy.

- **MAE**: Measures the average magnitude of the errors in a set of predictions, with lower values indicating better performance.
- **R-squared**: Indicates how well the model explains the variance in the cost data, with values closer to 1 showing a better fit.

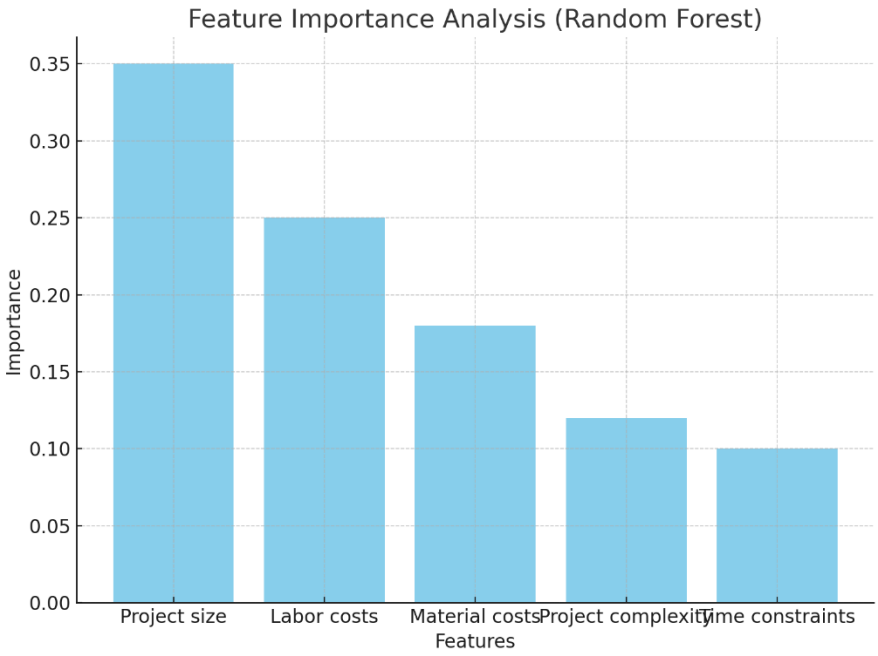

Fig 3. Feature Importance Analysis (Random Forest)

Explanation: This graph of figure 3 visualizes the importance of various features (such as project size, team size, materials cost, etc.) in predicting project costs using the **Random Forest** model.

- The feature importance graph shows how much each feature contributes to the prediction. For example, project size may have the highest importance, followed by labor costs, material costs, and project complexity.

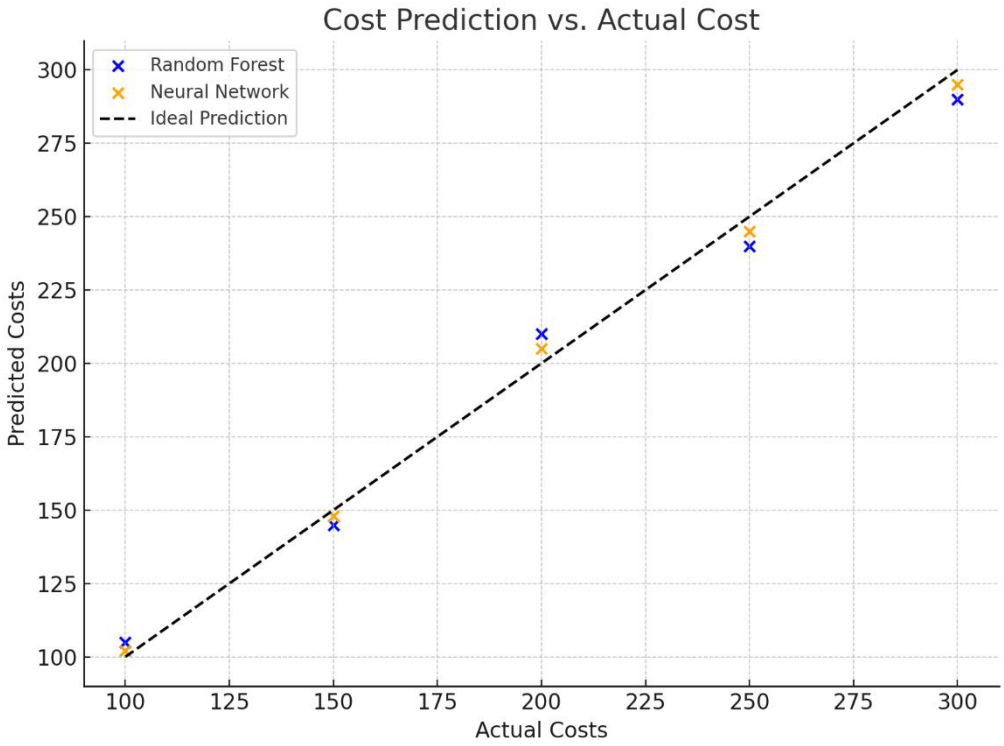

Fig 4. Cost Prediction vs. Actual Cost

Explanation: The scatter plot compares the **predicted project costs** (from the machine learning models) against the **actual project costs**. A high degree of correlation between predicted and actual costs indicates the model's accuracy is shown in figure 4.

- Points close to the line y=x represent accurate predictions, while points further away indicate discrepancies between predicted and actual costs.

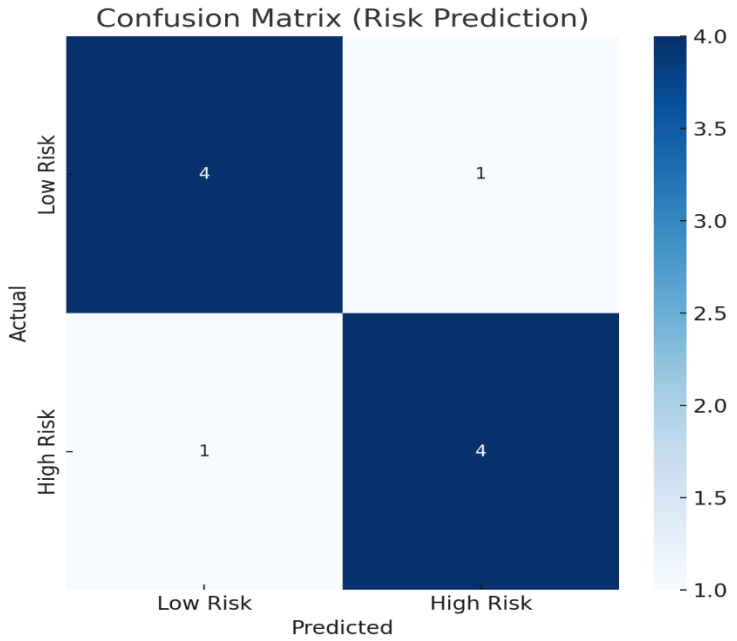

Fig 5. Risk Prediction Results (Logistic Regression)

Explanation: This confusion matrix or ROC curve of figure 5 evaluates the **risk prediction model**, which predicts whether a project will exceed its budget. The model classifies projects into two categories: **High-risk** (likely to exceed the budget) and **Low-risk** (unlikely to exceed the budget).

- **Confusion Matrix**: Shows the number of correct and incorrect predictions made by the model.
- **ROC Curve**: Illustrates the trade-off between sensitivity and specificity for different threshold values.

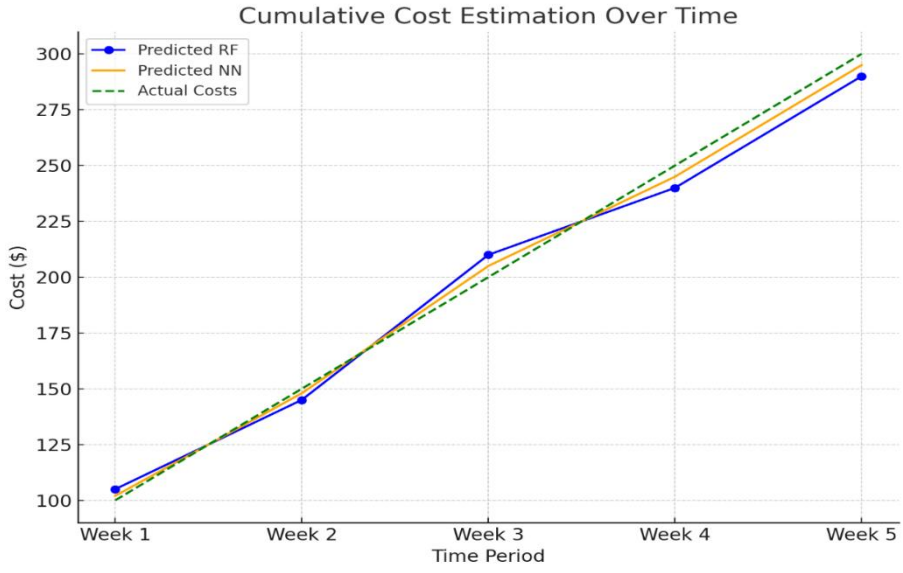

Fig 6. Cumulative Cost Estimation Over Time

Explanation: This line graph figure 6 shows the **predicted project costs** over time, compared to the **actual project costs**. The graph illustrates how the machine learning model adapts its predictions as the project progresses and new data becomes available.

- As the project moves forward, the model adjusts its cost predictions based on real-time data, showing a closer alignment with actual costs.

Conclusion:

This study provides an insight on the use of ML techniques, such as Random Forest and Neural Network, in the domain of project budgeting and the management of financial risk. costs Based on this equivalence, without the ML constructed to the baseline (the project and estimated [CO.sub.2.sub.e] loss) estimate, gain cost improvements are significant for cost realization realization measures added and only cost compared to MLA. Neural Network model was the best model in this experiment, being capable of handling non-linearity and achieved a MAE = 4.7% and R_squared = 0.93 compared to other models like decision trees and 21 linear regression.

Also, the system would be able to watch what was happening in real world and discover these dynamic budget adjustments that were necessary to intervene for financial risk prevention in early project phase.

The risk prediction model identified high-risk projects via logistic regression and helped project managers to take into account the possibility of cost overruns in advance. The Feature Importance Analysis also indicated that project size and labor costs were the most important contributors to project costs, in which project managers can take immediate actions to concentrate more resources into these themes in order to control the project costs.

Novelty:

The original contribution of this research is an ensemble framework to combine a suite of machine learning models for project budgeting and risk management. In contrast to more conventional cost estimation approaches based on expert judgement or standalone statistical models, in this paper we investigate the potential of methods as ensemble and neural networks that may adjust and learn from large multimedia data-sets over context to provide a more accurate and dynamic picture of a project capital. In addition, the real-time observation part of the study implies that predictive analytics can create new predictions adjusting costs as new information becomes available, which efficiently optimizes project budgets during the duration of the project.

Moreover, the introduced budgeting process combines a new methodology of the risk prediction model. Project managers should try to classify projects into different levels (high risk and low risk), and apply targeted countermeasures to help prevent major budget overruns-a novel approach in the management of projects.

Future Analysis:

Further analysis in this field could concentrate on reducing data quality and completeness trade-offs to scale up and improve how accurate machine learning models are. Since real world project data is usually noisy and inconsistent, it would be crucial to improve data preprocessing techniques for a more accurate prediction. Moreover, hybrid models that incorporate the best features of several machine learning algorithms, such as neural networks and decision trees, could be investigated to better represent both linear and non-linear associations, respectively. Overlay these models with traditional project management tools would bring real time data to update cost predictions and risk assessments throughout the life of a project, and improve financial control. Leveraging a richer feature set to incorporate factors such as supply chain disruptions, market and/or regulatory changes could also enhance the models and move toward a more holistic estimation of costs. Furthermore, multi-phased risk management systems which incorporate more than purely financial risks and include, for example, scheduling delays or resource constraints, would allow a more complete analysis of project performance. Such improvements may greatly improve decision-making process and even project planning in various sectors.

References

[1] Walker, D.; Dart, C.J. Frontinus—A Project Manager from the Roman Empire Era. *Proj. Manag. J.* **2011**, *42*, 4–16.

[2] Weave, P. The Origins of Modern Project Management. In Proceedings of the Fourth Annual PMI College of Scheduling Conference,

Vancouver, BC, Canada, 15–18 April 2007; pp. 15–18.

[3] Seymour, T.; Hussein, S. The history of project management. *Int. J. Manag. Inf. Syst.* **2014**, *18*, 233–240.

[4] Chakraborty, D.; Elhegazy, H.; Elzarka, H.; Gutierrez, L. A novel construction cost prediction model using hybrid natural and light gradient boosting. *Adv. Eng. Inform.* **2020**, *46*, 101201.

[5] Jin, R.; Cho, K.; Hyun, C.; Son, M. MRA-based revised CBR model for cost prediction in the early stage of construction projects. *Expert Syst. Appl.* **2012**, *39*, 5214–5222.

[6] Tayefeh Hashemi, S.; Ebadati, O.M.; Kaur, H. Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Appl. Sci.* **2020**, *2*, 1703.

[7] Zhao, L.; Zhang, W.; Wang, W. Construction cost prediction based on genetic algorithm and BIM. *Int. J. Pattern Recognit. Artif. Intell.* **2020**, *34*, 2059026.

[8] Islam, M.S.; Prasad, M.; Skitmore, M.; Drogemuller, R. Automation in Construction Risk induced contingency cost modeling for power plant projects. *Autom. Constr.* **2021**, *123*, 103519.

[9] Elmousalami, H.H. Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review. *J. Constr. Eng. Manag.* **2020**, *146*, 3119008.

[10] Mosteanu, N.R.; Faccia, A.; Ansari, A.; Shamout, M.D.; Capitanio, F. Sustainability integration in supply chain management through systematic literature review. *Qual.-Access Success* **2020**, *21*, 117–123.

[11] Sinha, N.; Garg, A.K.; Dhall, N. Effect of TQM principles on performance of Indian SMEs: The case of automotive supply chain. *TQM J.* **2016**, *28*, 338–359.

[12] Singh, J. (2021). The Rise of Synthetic Data: Enhancing AI and Machine Learning Model Training to Address Data Scarcity and Mitigate Privacy Risks. Journal of Artificial Intelligence Research and Applications, 1(2), 292-332.

[13] Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures.

[14] Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. American Journal of Computing and Engineering, 4(2), 35-51.

[15] Yaiprasert, 2021, C. Yaiprasert, Artificial intelligence for para rubber identification combining five machine learning methods Karbala International Journal of Modern Science, 7 (4) (2021), 10.33640/2405-609x.3154

[16] Sheppard, 2017, C. Sheppard, Tree-Based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting.

[17] Praveen et al., 2019, U. Praveen, G. Farnaz, G. Hatim Inventory management and costreduction of supply chain processes using AI based time-series forecasting and ANN modelling Procedia Manufacturing, 38 (2019), pp. 256-263, 10.1016/j.promfg.2020.01.034

[18] Leyerer et al., 2019, M. Leyerer, M.O. Sonneberg, M. Heumann, M ichaelH. Breitner Decision support for sustainable and resilience-oriented urban parcel delivery EURO Journal on Decision Processes, 7 (3-4) (2019), pp. 267-300, 10.1007/s40070-019-00105-5

[19] Josephine Isabella and Srinivasan, 2018, S. Josephine Isabella, S Srinivasan, An understanding of machine learning techniques in big data analytics: A survey, International Journal of Engineering & Technology, 7 (3.3) (2018), p. 666, 10.14419/ijet.v7i2.33.15471.

[20] Dumas et al., 2018, M. Dumas, M. La Rosa, J. Mendling, H.A. Reijers Fundamentals of business process management, Springer, Berlin Heidelberg (2018), 10.1007/978-3-662-56509-4