**A Comprehensive Analysis on Causes of Cost Overrun in Construction Projects Using Statistical and Machine Learning Methods**

1Pramodini Sahu, 2Ayush Nanda

1Assistant Professor, Department of Civil Engineering, Odisha University of Technology and Research, Email: [pramodinice@outr.ac.in](mailto:pramodinice@outr.ac.in)

2 Department of Information Technology, Odisha University of Technology and Research, Email: [ayushnanda.3110@gmail.com](mailto:ayushnanda.3110@gmail.com)

**Abstract**

Cost overruns in construction projects are considered one of the most recurrent problems affecting the financial and operational output of the projects. This research incorporates statistical and machine learning approaches to analyse the casual factors behind the cost overruns in construction projects. The present study applies an integrated data set that reflects other project characteristics such as project size, complexity, delays resource allocation and reasons relevant to the escalation of cost overruns. Resources allocation, and reasons relevant to the escalation of cost overruns Patterns and relationships between project variables and cost overruns are determined by the application of statistical method using Relative Important Index (RII). Moreover prediction of cost overrun is possible by the use of a machine learning algorithms. This research focused on Random Forest Genetic Algorithm (RFGA) , based on the project attributes. The combination of these approaches allows for a more robust and accurate identification of factors influencing cost overruns, offering insights that traditional methods may overlook. The findings of this study not only contribute to the academic understanding of cost overruns but also provide practical recommendations for project managers and stakeholders in the construction industry to mitigate risks and improve cost control strategies. Advanced data-driven techniques help this study draw attention to the importance of proactive decision making and resource management in reducing construction cost overruns. Cost overruns in construction projects are considered one of the most recurrent problems affecting the financial and operational output of the projects. This research incorporates statistical and machine learning approaches to analyse the casual factors behind the cost overruns in construction projects.

**Introduction:**

The construction project performance is generally expressed in terms of time and cost variance against its baseline. Out of the four fundamental constraints namely scope, cost, time and quality, cost performance is the most essential and common issue in the global construction industry. It is important to measure the cost variance in construction to understand the performance of the project and thereby to understand financial risks involved in the project execution. The cost variance, resulting as project cost overrun is denoted as a negative impact on economy and the profitability. Several perspectives on cost overrun are available in extant literature and are proven valid. However, it is not adequately explained on why the cost overrun keeps occurring though sufficient knowledge on cost overrun has been largely shared. The causes of inaccuracy in forecasts are different for different projects. The change of governance structures for forecasting the project development as an external factor also threat the project planning and execution. It was found that projects do not perform as predicted, in terms of costs: almost 9 out of 10 projects fall victim to significant cost overrun. Cost overrun is a persistent problem in construction projects worldwide. Understanding the causes of cost overrun is critical for project managers, engineers, and stakeholders to ensure more effective budgeting, forecasting, and risk management strategies

**Literature review:**

Numerous studies have attempted to identify the root causes of cost overruns, categorizing them into various factors such as poor project management, inaccurate cost estimations, unexpected site conditions, and changes in project scope. According to Flyvbjerg et al. (2003), large infrastructure projects are particularly susceptible to cost overruns, with an average cost escalation of 28%. These overruns are attributed to a combination of technical, economic, and psychological factors that complicate project planning and execution. Traditional methods of analysing cost overruns have predominantly relied on statistical techniques, which, while valuable, often fail to capture the complex, non-linear interactions between various factors. These methods generally assume a linear relationship between independent variables and the dependent variable (cost overrun), potentially oversimplifying the problem. Moreover, traditional statistical models may not adequately handle the high dimensionality and heterogeneity of construction data, leading to limited predictive accuracy (Love et al., 2011). Studies have consistently emphasized the multifaceted nature of cost overruns. The study by Kaming et al. (1997) identified design changes, inaccurate estimates, and project complexity as major causes of cost overruns in large construction projects while Aibinu and Jagboro (2002) focused on the role of costs the excess dominates, which means that project delays tend to increase costs due to project length and inflationary pressure. External factors such as economic conditions and regulatory changes also play an important role. According to Memon et al. (2012), external factors such as commodity price fluctuations, labour shortages and political instability can severely constrain project budgets. These findings emphasize the importance of comprehensive risk management strategies to mitigate externalities. Economic factors, including fluctuations in material prices, labour costs, and inflation, also play a crucial role in cost overruns. Construction projects are particularly sensitive to economic volatility, which can lead to unanticipated budget increases. Furthermore, external factors such as regulatory changes, environmental conditions, and political instability can disrupt project timelines and budgets, exacerbating the issue of cost overruns. Traditional statistical methods, such as regression analysis and correlation studies, have become the cornerstone of cost-over-analysis. However, these methods often fail to accommodate the complex and dynamic nature of construction. For example, Love et al. (2005) pointed out that statistical models oversimplify the complex interactions between different elements of the project, leading to limited predictive power. Furthermore, statistical methods typically assume that the relationship between variables is linear, which may not be true in multi-stage construction projects. Despite their limitations, the statistical methods have provided a starting point for understanding the causes of cost overruns. For example, Kaushki et al. (2005) used regression analysis to identify key predictors of cost overruns in residential construction, including project size, contractor experience, and project financing strategies.The upward thrust of machine learning (ML) has opened new frontiers in price overrun prediction. ML models, including assist vector machines (SVM), random forests, and neural networks, offer superior abilities for studying massive and complicated datasets. These fashions can find non-linear relationships and interactions amongst challenge variables, main to more accurate and insightful predictions. Research with the aid of Jarkas and Horner (2015) proven the effectiveness of ML algorithms in predicting construction venture outcomes. Their study observed that ML models drastically outperformed conventional statistical methods in terms of predictive accuracy and robustness. Similarly, Kim et al. (2013) carried out synthetic neural networks to predict cost overruns in constructing construction projects, achieving better accuracy compared to traditional techniques. The adaptability of ML models is some other giant gain. As new facts turns into to be had, those models may be retrained to enhance their predictive accuracy, making them surprisingly dynamic tools for real-time decision-making (Shahin et al., 2019).Combining conventional statistical strategies with gadget mastering gives a holistic approach to information and predicting fee overruns. Statistical strategies are beneficial for preliminary records exploration and figuring out key variables, whilst ML fashions excel in shooting complicated patterns and interactions. This included approach has been recommended by numerous researchers to decorate the general predictive overall performance and reliability of price overrun fashions. For instance, Cheng et al. (2010) proposed a hybrid model that combines regression evaluation with neural networks to predict price overruns in production projects. They found that the hybrid model provided superior predictive accuracy as compared to standalone statistical or ML models. This included technique no longer handiest improves predictive accuracy however additionally gives a deeper knowledge of the underlying causes of value overruns.

There are many research gaps, despite a broad literature that addresses cost overruns in construction projects. Traditional statistical methods, while offering foundational insights into the causes of cost overruns, often neglect the complex and non-linear assumptions are normally taken for such analysis, but suffer from dimensional difficulties when dealing with high-dimensional data in addition to the heterogeneity encountered in construction projects. Many studies have focused on specific project types or regions, which limits the generalizability of their findings. Further research is needed to explore the global applicability of integrated statistical and ML methods, considering diverse project characteristics and environmental factors. This comprehensive approach can help bridge the existing gaps and provide more robust solutions for mitigating cost overruns in construction projects.

**Objective:**

Cost overrun may lead to financial losses for all the stakeholders. Cost overrun may be the result of various unforeseen causes which may lead to conflicts between contractors, owners and various involved parties. By identifying the root causes and analysing them can provide us a better understanding on how these causes contributes to disputes and financial strain. In this analysis, we aim to identify the key factors contributing to cost overruns and the level was ranked considering its Relative Important Index value calculated from traditional statistics. Also the research uses Random Forest Genetic Algorithm (RFGA) model to make prediction of the riskiness level.

**Methodology:**

A well-defined questionnaire was prepared with consultation of stakeholders of construction projects. The questionnaire had two sections. The first section included the details of the project and the second part sought information for causes of cost overruns as given in Table-1, and respondents were requested to provide the severity of cause with respect to 5-point Likert scale (Very low, Low, Moderate, High Extreme weighing 1 to 5). Data were collected in a pilot survey for different types of construction projects in India.

Table-1: Causes of cost overruns construction project

|  |  |
| --- | --- |
| Cause No | Causes of Cost Overrun |
| C-1 | Improper cash inflow and outflow |
| C-2 | Errors in designs |
| C-3 | Delays in analysis & alteration in designs |
| C-4 | Mismatch of ongoing construction with the actual plan |
| C-5 | Increase in the cost of project |
| C-6 | Conflict between owner and contractor |
| C-7 | Changes in soil characteristics, at the time of construction |
| C-8 | Inadequate risk management |
| C-9 | If owner is not financially viable |
| C-10 | Issues relating to frauds and miscreants |
| C-11 | Poor construction strategy |
| C-12 | Poor site management |
| C-13 | Enhancement in the cost of labourers |
| C-14 | Unusual payment to labourers |
| C-15 | Improper control of finance |
| C-16 | Improper communication between owner and the contractor |
| C-17 | Conflict on differentiating and biasing in between labourers |
| C-18 | Improper material management and procurement |
| C-19 | Accident due to faulty equipment |
| C-20 | Non-availability of authorized (IS marked) machineries |
| C-21 | Sudden fall of demand of a particular material in the market |
| C-22 | Delivery of faulty machineries |
| C-23 | Idle mindset of labourers |
| C-24 | Improper maintenance of equipment |
| C-25 | Changes in materials and labour |
| C-26 | Demand of 4 prices by group of labourers |
| C-27 | Poor project planning and estimation |
| C-28 | Failure of equipment at the time of construction |
| C-29 | Price hike of materials |
| C-30 | Hiring of unskilled labourers |
| C-31 | Improper technical performance |
| C-32 | Lack of technical assistance from the government |
| C-33 | Wastes rising from construction materials |
| C-34 | Existing demand in market |
| C-35 | Less availability of living and food arrangements for labors. |
| C-36 | Inadequate materials |
| C-37 | Improper use of resources |
| C-38 | Worker’s problems |
| C-39 | Sudden increase in price of commodities |
| C-40 | Less availability of time (time uncertainties). |
| C-41 | Inaccurate project cost, which leads to confusion. |
| C-42 | New additional works, which gave rise to increased problems |
| C-43 | Problems pertaining to environmental clearance |
| C-44 | Unpredictable circumstances (flood, cyclone, earthquake etc) |
| C-45 | COVID-19 pandemic |
| C-46 | Levies or taxes charged by the agencies |
| C-47 | Local people/political interference |
| C-48 | Unsuitable weather conditions |

The ranking of the causes of delay in construction projects was done concerning its RII value in conventional statistical method. Similar methods also adopted by Gunduz et al (2013), Sambasivan and Soon (2007), and Kometa et al. (1994) for analysis of causes of delay in construction sectors.

Relative Important Index (RII) is based on the severity



a - weightage given to each response (from 1 for very low to 5 for extreme) n -frequency of the responses N- Total number of responses

***Prediction of risk of cost overrun using Random Forest Genetic Algorithm (RFGA) model:***

RFGA model was used to foresee the risk of cost overrun. The RFGA model is coded using the Jupytor Notebook. The input variables were all the cost overrun causes and the model used the severity of causes’ data set. The total data was catagorised into two phases: training data 30% and testing data 70% in the model and genetic algorithm is applied for optimization. A similar model was adopted by Yaseen et al (2020) to forecast the risk in construction delays.

**Result and Discussion:**

***Characteristics of respondents:***

The questionnaire was sent to more than 100 people including contractors, owners, consultants, field engineers, material suppliers, project managers, surveyors, etc. The survey was made on different government projects as well as private projects including different type of projects such as building project, bridge and dam projects, highway projects, maintenance works, pipe line and drainage works etc. The data analysis of the current research work has been conducted for 32 projects.

***Ranking of causes with respect to RII:***

The RII value was calculated in percentage and the higher the RII percentage indicates the severity of causes towards cost overrun.Table-2 illustrates factors like price hike of materials, increase in the cost of project, poor project planning and estimation, problems pertaining to environmental clearance and COVID-19 pandemic, inadequate risk management, time uncertainties lead to high risk of cost overrun in construction projects. The below table shows the RII value of each factor that causes financial loss to the projects.

Table-2: Ranking of causes with respect to RII

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cause No** | **RII(%)** | **RII ranking** |  | **Cause No** | **RII(%)** | **RII ranking** |  | **Cause No** | **RII(%)** | **RII ranking** |
| C-29 | 69.38 | 1 |  | C-15 | 55.63 | 17 |  | C-11 | 51.25 | 33 |
| C-5 | 63.75 | 2 |  | C-19 | 55.63 | 18 |  | C-13 | 51.25 | 34 |
| C-27 | 63.75 | 3 |  | C-4 | 53.75 | 19 |  | C-18 | 51.25 | 35 |
| C-43 | 63.75 | 4 |  | C-16 | 53.75 | 20 |  | C-47 | 51.25 | 36 |
| C-45 | 62.50 | 5 |  | C-17 | 53.75 | 21 |  | C-2 | 50.63 | 37 |
| C-39 | 61.88 | 6 |  | C-22 | 53.75 | 22 |  | C-24 | 50.00 | 38 |
| C-41 | 61.88 | 7 |  | C-26 | 53.75 | 23 |  | C-25 | 50.00 | 39 |
| C-40 | 61.25 | 8 |  | C-30 | 53.75 | 24 |  | C-31 | 49.38 | 40 |
| C-8 | 60.63 | 9 |  | C-9 | 53.13 | 25 |  | C-35 | 49.38 | 41 |
| C-21 | 60.00 | 10 |  | C-20 | 53.13 | 26 |  | C-23 | 48.13 | 42 |
| C-1 | 58.75 | 11 |  | C-36 | 53.13 | 27 |  | C-32 | 47.50 | 43 |
| C-3 | 58.13 | 12 |  | C-44 | 53.13 | 28 |  | C-12 | 46.25 | 44 |
| C-7 | 57.50 | 13 |  | C-48 | 52.50 | 29 |  | C-33 | 45.00 | 45 |
| C-28 | 57.50 | 14 |  | C-14 | 51.88 | 30 |  | C-38 | 42.50 | 46 |
| C-37 | 57.50 | 15 |  | C-34 | 51.88 | 31 |  | C-46 | 41.88 | 47 |
| C-6 | 55.63 | 16 |  | C-42 | 51.88 | 32 |  | C-10 | 36.88 | 48 |

***Risk prediction:***

Figure-1 represents the risk levels predicted by the RFGA model. It illustrates the causes such as Delays in analysis & alteration in designs, COVID-19 pandemic, Less availability of time (time uncertainties), Improper control of finance, Inaccurate project cost, which leads to confusion and Accident due to faulty equipment are at high risk of becoming a conflict.

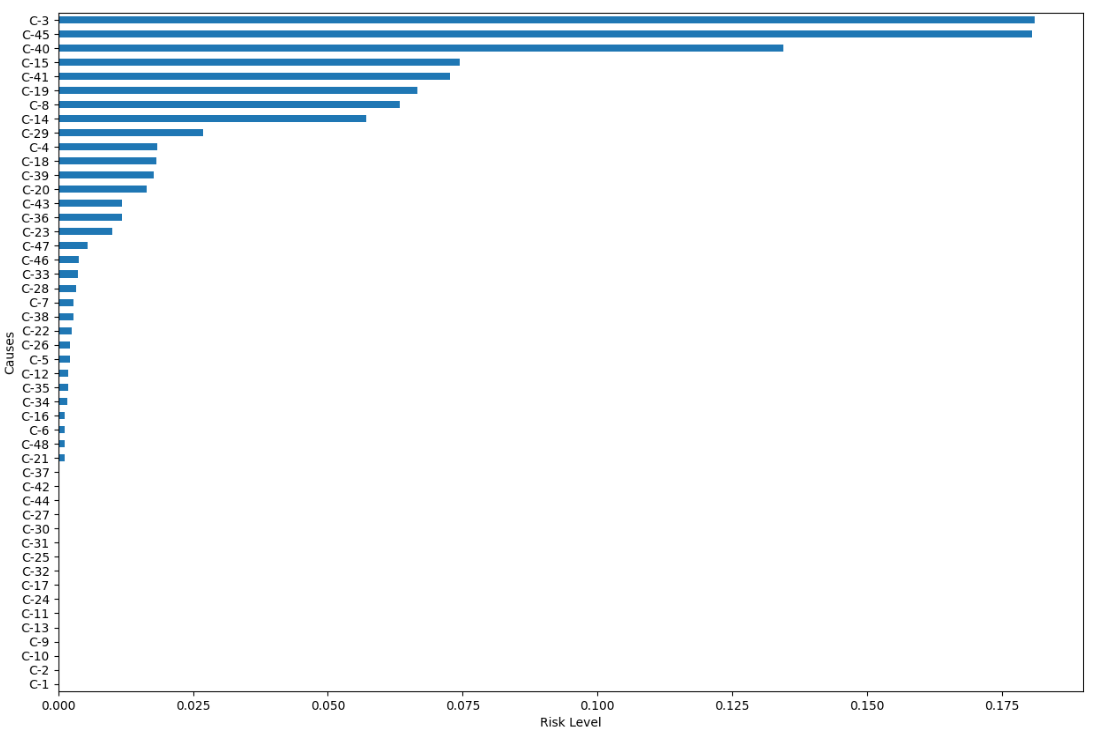


Figure-1: Barchart showing risk level of causes of Cost overrun using RFGA model

The statistical performance of the training and testing data sets of the RFGA were assessed using the model performance measures derived from the confusion matrix from Figure-2 and Figure-3.

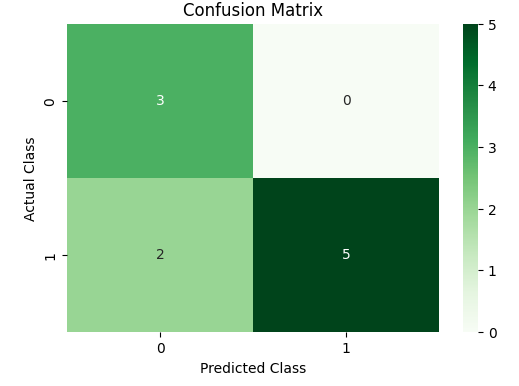
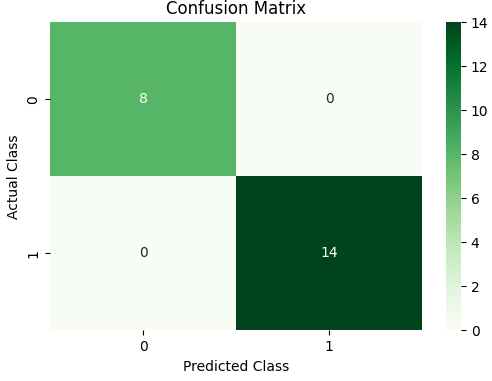


Figure-2: Confusion matrix(Training data)Figure-3: Confusion matrix (Testing Data)

In Table 5 performance of training and test data is predicted using precision recall and accuracy. The accuracy of the RFGA classifier in training data is 1.00 and the weighted precision in the category of 1(Positive cost overrun) is 1 whereas in 0(no cost overrun) group is 1 similarly recall values are found to be 1 and 1 respectively. The accuracy of the testing data is 0.80, weighted precision and recall values are 0.6 and 1.0 in categories of 1(Positive cost overrun) and 1 and 0.71 for category of 0(no cost overrun) respectively.

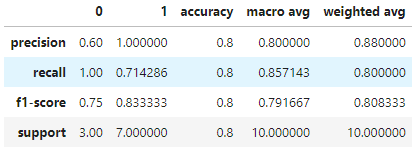
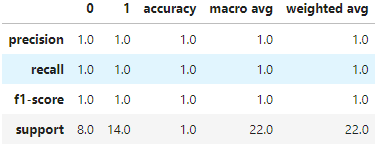


Figure-4: Performance measures in RFGA classifier for Training data and Testing Data

The model predicted major causes that are at high risk towards cost overrun. The steps that can be taken to overcome are good choices of materials, laborers, have a good attitude and strategy towards the project and should try the best to complete the project within a given time, irrespective of conditions; the supervisors must check the progress of the project at regular intervals. At last Proper coordination between the owner and structural consultant is crucial.

**Conclusion:**

A detailed analysis of the gathered data on 32 construction projects was conducted to determine the actual reason of cost overrun claims in the construction industry, as well as their relative important index in terms of incidence, using a statistical method and the RFGA machine learning model. The data were accumulated on a 1 to 5 scale for causes of Cost overrun claims concerning their occurrence throughout the project. Additionally, the responders were asked to mention the possibility of cost overrun. The RII value was calculated for each cause. The cause with greater RII value indicates more severe towards cost overrun in the project. Cost overrun causes like price hike of materials, increase in the cost of project, poor project planning and estimation, problems pertaining to environmental clearance and COVID-19 pandemic, inadequate risk management, time uncertainties are the most critical causes. To predict the dispute from delay claims RFGA approach was used to categorize project delays in two groups as 1(positive overrun) and 0(no overrun) based on respondent data and schedule overrun percentage. The model had accuracy of 1.00 and 0.80 for testing and training data respectively. The model predicted that factors like Delays in analysis & alteration in designs, COVID-19 pandemic, Less availability of time (time uncertainties), Improper control of finance, Inaccurate project cost, which leads to confusion and Accident due to faulty equipment are at high risk of becoming a conflict. The current study will improve economic impact, construction industrial reputation, and project management.

**References**

1. Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects?. *Transport reviews*, *23*(1), 71-88.
2. Love, P. E., & Edwards, D. J. (2004). Determinants of rework in building construction projects. *Engineering, Construction and Architectural Management*, *11*(4), 259-274.
3. Kaming, P. F., Olomolaiye, P. O., Holt, G. D., & Harris, F. C. (1997). Factors influencing construction time and cost overruns on high-rise projects in Indonesia. *Construction Management & Economics*, *15*(1), 83-94.
4. Aibinu, A. A., & Jagboro, G. O. (2002). The effects of construction delays on project delivery in Nigerian construction industry. *International journal of project management*, *20*(8), 593-599.
5. Memon, A. H., Rahman, I. A., & Azis, A. A. (2012). Time and cost performance in construction projects in southern and central regions of Peninsular Malaysia. *International Journal of advances in applied sciences*, *1*(1), 45-52.
6. Jarkas, A. M., & Horner, M. W. (2015). Creating a baseline for labour productivity of reinforced concrete construction using systematic site observations. Construction Management and Economics, 33(3), 246-260.
7. Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and environment*, *39*(10), 1235-1242.
8. Koushki, P. A., Al‐Rashid, K., & Kartam, N. (2005). Delays and cost increases in the construction of private residential projects in Kuwait. *Construction management and economics*, *23*(3), 285-294.
9. Shahin, M. A., Jaksa, M. B., & Maier, H. R. (2001). Artificial neural network applications in geotechnical engineering. *Australian geomechanics*, *36*(1), 49-62.
10. Cheng, M. Y., Tsai, H. C., & Hsieh, W. S. (2009). Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model. *Automation in construction*, *18*(2), 164-172.
11. Gündüz, Murat, Yasemin Nielsen, and Mustafa Özdemir. "Quantification of delay factors using the relative importance index method for construction projects in Turkey." Journal of management in engineering 29.2 (2013): 133-139. 10.1061/(ASCE)ME.1943-5479.0000129
12. Horta, I. M., et al. "Performance trends in the construction industry worldwide: an overview of the turn of the century." Journal of productivity analysis 39 (2013): 89-99. DOI 10.1007/s11123-012-0276-0
13. Kometa, Simon T., Paul O. Olomolaiye, and Frank C. Harris. "Attributes of UK construction clients influencing project consultants' performance." Construction Management and economics 12.5 (1994): 433-443. 10.1080/01446199400000053
14. Yaseen, Zaher Mundher, et al. "Prediction of risk delay in construction projects using a hybrid artificial intelligence model." Sustainability 12.4 (2020): 1514. 10.3390/su12041514