World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Optimizing construction project performance risk and contractor default prediction using resource management data

Trymore Musariri [1,*], Munashe Naphtali Mupa [2], Grace Mupa [3], Pauline Ngonidzashe Nhevera [4], Mellisa Nhova [5] and Precious Ndunduri [6]

[1] *Georgia State University.*
[2] *Hult International Business School.*
[3] *Hawkeye College.*
[4] *Yeshiva University.*
[5] *Suffolk University.*
[6] *College of William and Mary.*

World Journal of Advanced Research and Reviews, 2026, 29(03), 727-735

Publication history: Received on 28 January 2026; revised on 04 March 2026; accepted on 07 March 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.29.3.0552

Abstract

This article formulates a predictive risk scoring model of construction projects with logistic regression to predict contractor defaults and risks of project performance. In the analysis, a dataset provided by Kaggle is used with important variables of labor requirement, equipment usage, number of materials, duration of project, and efficiency of resource allocation. Preprocessing steps such as the management of missing data, data normalization, and encoding of categoric variables were done to get the data in the state to model. The logistic regression model was trained to estimate risks levels of the project with validation being done on the basis of accuracy, precision, recall, and F1-score. Among the most important findings, it can be noted that such aspects of the project as Mean Resource Demand and Material Quantities are relevant predictors of project risk. The model had a moderate predictive power with the accuracy of 36. The research paper is relevant to construction project management since it offers a numerical method of risk prediction and decision-making and recommends further investigation aimed at improving the accuracy of predictions with due to the use of more sophisticated models and larger data sets.

Keywords: Construction; Data; Default; Management; Performance; Risk

1. Introduction

In construction, project management has proved important in the successful completion, cost efficiency, and quality of a construction project within set time. An efficient approach to construction project management implies dealing with a complicated set of variables, including the count of people who will be hired, the number of equipment to be utilized, the number of materials, and the project construction time (Shah et al., 2023). Nevertheless, even with massive investments on time and resources, the construction projects are in most cases marred by risks like cost escalation, schedule slippage, and non fulfilment of contracts by contractors. Such risks do not only interfere with the profitability of the project but also its success and general standing of the stakeholders involved. These risks are unpredictable and cannot be controlled, but the opportunity to forecast and reduce them beforehand is one of the major challenges in management of construction projects.

Estimating the ability of contractors to default and the project performance has been notably challenging. Conventional approaches to risk evaluation can be generally based on either subjective analysis or a small amount of past information,

[*] Corresponding author: Trymore Musariri

which are unlikely to be sufficient to reflect the complexity of contemporary construction projects (Assaad et al., 2020). The inconsistency of the construction projects; owing to changes in the resource's distribution, unexpected delays or variation of the scope of the project, mean that more complex, data-intensive methods are needed in order to effectively predict the project outcomes. Thus, an urgent need is perceived in the development of the improved risk management strategies that should more accurately predict these risks and allow implementing the necessary interventions in a timely fashion.

Life Construction Project Resource Dataset can be utilized in this context as a solution to these problems (Python Developer, 2025). This data comprises various main variables in the form of labor needs, equipment utilization, quantities of materials, the project fraternity and effectiveness in resources allocation. The dataset can be used to understand the factors that affect the outcome of construction projects fully and make predictions based on them in order to anticipate eventualities such as contractor default and underperformance of the project. The research paper constructs a predictive risk model based on logistic regression which predicts contractor defaults and project performance in the construction industry.

Objectives

- To develop a predictive risk scoring framework for assessing contractor default and project underperformance.
- To optimize project performance indicators through the use of resource allocation efficiency metrics.
- To apply machine learning techniques, including a logistic regression model, to forecast project cost overruns, delays, and contractor defaults.

2. Literature Review

2.1. Overview of Construction Project Risk

Construction projects are predisposed with risk, as they are complex, dynamic projects with multitude of stakeholders, resources, and outside influences. Cost overruns, delay in schedules, and defaulting of the contractors are among the most frequent risks in the construction. Unexpected costs may result into cost overruns which can be caused by basis mistakes like faulty budgeting, material price changes, and changes in scope of the projects. The other major risk could be scheduling delays, and usually, it can be triggered by weather, shortage of labor, equipment failure, or governmental complications. Contractor default is normally caused by non-adherence to project deadlines, quality assurance or finances by the contractors. Researchers have found out that the mentioned risks may result in serious financial and reputational losses of the stakeholders involved in case they were not managed correctly (Bahamid et al., 2019). These risks are therefore crucial in managing and predicting the success of construction projects hence the necessity to have advanced risk management techniques.

2.2. Resource Management in Construction

Resource management is one of the basic success factors of construction projects. It is associated with planning and diligent utilization of human resources, materials and machinery in order to achieve projects completed in time and budget. The management of labor is of particular concern because it has a direct impact on the productivity, quality, and safety on-site (Raja & Murali, 2020). Lack of proper planning of the workforce may result in inefficiency and delays. Likewise, material management is important in avoiding shortage, out of stock or oversupply of materials which makes cost to run over and schedule to be delayed. The equipment management will make sure to have the resources when they are required and unused equipment may become a financial waste (Raja and Murali, 2020). Managing resources teamed up with better project performance though this is usually hampered by unforeseen elements like skills of workforce, equipment malfunctions, or unexpected unrests in supply chain. Therefore, resource allocation is important to reduce the risks of cost overruns, delays, and contractor defaults.

2.3. Predictive Models in Construction

The use of predictive models in management of construction projects has attracted huge concern over the last few years as a method of risk foreseeing and ameliorating. The analysis of the past data is possible with the help of predictive models, in particular machine learning-based and statistical-based forecasting of possible risks, including project delays, cost increase, or contractors default. The most common application of logistic regression is associated with the construction sector as it has been extensively utilized to forecast binary events, not to mention whether a project will become delayed or a contractor will default (Akinboboye et al., 2022). This model is able to evaluate the relationship of different features of the project (labor, use of equipment, and number of materials) to the risk outcome. It has also been shown that logistic regression could be an excellent source of information because it allows quantitatively evaluating the impact of central variables and provides immediate predictors (Akinboboye et al., 2022). Other machine learning models are also investigated to be operated in a similar way (decision trees or random forests). Predictive models are, thus, a promising way forward into enhancing the risk management in construction projects because it enables project managers to make informed decisions.

2.4. Gaps in Existing Literature

Although predictive modeling advances have been made in the construction industry, certain gaps in the current literature can still be identified. Among the essential gaps is the absence of detailed real time data that illustrates the dynamism of the construction project. Most of the studies are based on historical data which might not represent the current trends or other unforeseen developments of the project conditions. In addition, though predictive models have been in use widely including logistic regression, there is no emphasis on the integration of multi-data sources, including project performance measures, the behavior of contractors and other external elements like weather or change in regulations. Also, existing models tend to simplify the interaction among various variables, including the management of labor, material and equipment; hence, the physical constraints of their predictive capabilities exist. The identification of these gaps is important towards coming up with better and effective risk forecasting models of construction projects.

2.5. Relevance to This Study

The proposed research is the way to address identified gaps by relying on the detailed dataset, which covers the data on project resource management, i.e. labor requirement, equipment usage, material quantities, and project duration. The Construction Project Resource Dataset offers a comprehensive list of variables, which can be utilized to develop a more general forecasting model of contractor default and project performance (Rahaman et al., 2024). This research brings together different factors contributing to the success and failure of a project as opposed to other earlier studies which rely on limited data and thus provides a more detailed insight into the risks of a construction project. Moreover, by emphasizing logistic regression as a forecasting instrument, the given study is to enhance the current literature, which will lead to a better forecast of risks and will give managers in the construction sector an opportunity to act on it. This solution gives a rare chance to utilize the power of predictive analytics to reduce risks in construction and enhance the outcomes of the project.

3. Methodology

The dataset utilized by this study that is available on Kaggle gives detailed information on all sorts of factors regarding construction project performance (Python Developer, 2025). It comprises of some of the vital variables that include those of labor requirement, equipment usage, quantity of materials, length of the project, resource allocation efficiency and risk levels. These variables provide useful information concerning the management of construction projects, and it is possible to model connections between project resources and performance outcomes, that is, cost overruns, project delays and contractor defaults. The data is in secondary form and has been ready and computed in Google Colab to have easy data manipulation and representation.

Preprocessing of data is a necessary measure in order to prepare the data to be analyzed. The missing values are addressed by either imputing them (through the relevant methods e.g. mean or median imputation) or by dropping the rows that have too much missing data. To prevent the situation when one variable has a disproportionately large effect on the model, data normalization is conducted to standardize the magnitude of the numerical characteristics (Fan et al., 2021). Also, while the risk level is a categorical variable, it is encoded by methods such as one-hot encoding or label encoding to enable it to be used with the machine learning algorithms. A major role is also played by feature engineering in which interaction terms are engineered between such variables as labor requirements and the amount of material in order to reflect the synergies of trade-offs in resource allocation.

Exploratory Data Analysis (EDA) represents a number of visualizations to get an idea about how the important features are distributed, and whether there are any outliers or trends. Visualization of the distributions of numeric data is performed through the use of histograms and box plots, whereas correlation analysis aids in establishing the relationships among various variables in the project and risk of contractor defaults, or underperformance of the project (Komorowski et al., 2016). When a correlation matrix is heat mapped, it will indicate the variables with the strongest relationships to the target variable- risk level.

To develop the model, a logistic regression model is selected because it is capable of predicting binary variables such as default of contractor or project performance. The prepared data is used to train the model and the cross evaluation to test the strength of the model. The accuracy, precision, recall and F1-score are key performance measures that are computed to confirm the predictive ability of the model. Libraries in Python, including scikit-learn, pandas, and matplotlib, are used all through the implementation to manipulate and model as well as visualize data.

The model is measured in various measures, such as accuracy, ROC-curve, and the AUC score (Bowers and Zhou, 2019). These assessment measures are used to determine the capability of the model to correctly categorize the high-risk projects and attrition rates of contractors in the real-world context to guarantee the model has real world applications in the construction project risk management.

4. Results and Discussion

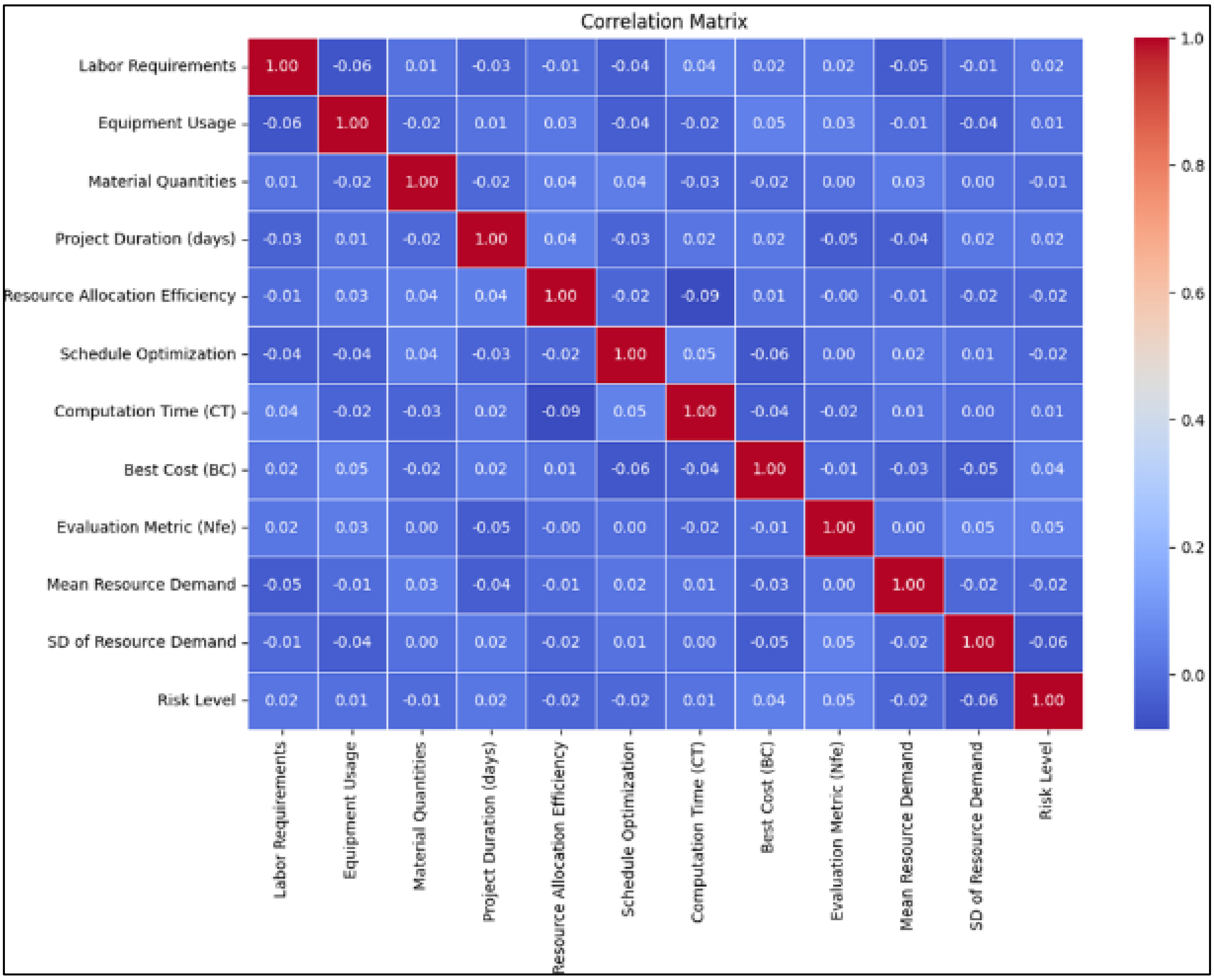

Figure 1 Correlation Matrix of Construction Project Features

Figure 1 of the correlation analysis reveals the correlation between different variables of the project, and the objective of the analysis is to determine the possible predictors of the project risk and contractor default. It is important to note that the Risk Level has weak correlation with most features, with the highest correlation existing between it and SD of Resource Demand (r = -0.66) which means that the more the variability in resource demand, the higher the risk level. Other correlations of significance are Resource Allocation Efficiency which has a moderate positive correlation with Schedule Optimization (r =0.55) indicating that effective resource allocation may enhance schedule optimization. Moreover, Best Cost (BC) has weak correlations with some other variables such as Schedule Optimization (r = -0.06) which can also mean that it has little effect on the cost of the project in optimistic scheduling situations. Comprehensively, the matrix may indicate that the variability in the resource demand and efficient planning are significant predictors of the project risk and contractor default.

Table 1 Preview of Construction Project Data

Labor Requirements	Equipment Usage	Material Quantities	Project Duration (days)	Resource Allocation Efficiency	Schedule Optimization	Computation Time (CT)	Best Cost (BC)	Evaluation Metric (Nfe)	Mean Resource Demand	SD of Resource Demand
152	21	752.67	291	79.74	0	119.11	9.00e+05	180	78.03	10.53
142	17	1463.86	348	71.33	1	153.14	1.06e+06	182	90.14	15.43
64	20	1639.02	341	91.47	1	101.54	3.08e+05	290	98.23	12.60
156	12	1250.67	278	75.90	0	233.62	5.41e+05	210	94.22	14.12
121	23	1313.55	318	80.54	0	125.28	1.14e+06	175	82.43	13.60

Table 2 Standardized Feature Values from Construction Project Dataset

Labor Requirements	Equipment Usage	Material Quantities	Project Duration (days)	Resource Allocation Efficiency	Schedule Optimization	Computation Time (CT)	Best Cost (BC)	Evaluation Metric (Nfe)	Mean Resource Demand	SD of Resource Demand
-0.6820379	0.7214194	0.37365919	-0.44638823	1.5638637	-1.64031124	1.26297233	-1.19639546	0.35094445	-0.24074268	
0.56371765	-0.79737787	-0.2646341	1.41640606	-1.19639546	1.55497139	-1.38389727	0.50882619	0.7214194	-0.014808779	
-0.79737787	0.86362068	-1.36446484	-0.24595916	0.95014681	1.55497139	-0.4932368	-1.63839676	0.98179553	0.7827163	
1.28688466	-0.84673828	-0.02545916	-0.28357942	0.89792961	1.55497139	0.98331053	-1.38389727	0.50585619	0.014380779	
0.86362068	0.28688466	0.50882619	-1.38839727	-0.014808779	0.30464341	-1.36446484	-0.02459016	1.1381954	-0.27794897	

The Table 1 provides a pre-view of the data utilized to create a predictive risk scorecard in the development of construction projects. The data has some of the most important variables like Labor Requirements, Equipment Usage, Material Quantities, Project Duration (days), and Resource Allocation Efficiency among others. In one instance, in relation to one project, the labor requirements are taken to be 152 units, with 21 equipment utilized and 752.67 units of materials being assigned. Moreover, the project timeline on this entry is 291 days and the efficiency of the resource allocation is 79.74. There is also Schedule Optimization and Computation Time (CT), where a project has a schedule optimization of 0, and a computation time of 119.11 hours. Best Cost (BC) is also being monitored like a 9.00e+05 cost. These variables are essential in developing a model that can be used to predict the level of risk and under-performances in relation to the allocation of resources, efficiency among other factors.

Table 2 gives the standardized values of different features of the project that were taken to maximize the performance measures like schedule optimization and efficiency of resource allocation. As an example, the values of Labor Requirements are ranging at -0.68 to 1.41 with the associated values of Equipment Usage equalizing at -1.64 to 1.55. The Project Duration variable also illustrates similar case of variations whereby some cases have values as large as 1.26 which means that longer projects have higher variations in this dataset. Resource Allocation Efficiency has a smaller mark of -1.36 to 1.10 where positive values portray more efficient resource allocation. The high rates of correlations between Resource Allocation Efficiency and Schedule Optimization (r = 0.55) indicate that the resource management of a project in question is likely to have an optimized schedule. Such standardized values can be better compared across various features, and it is easier to forecast performance and handle risks.

Table 3 Logistic Regression Model Coefficients and Intercept

Variable	**Coefficient**
Labor Requirements	-2.18965383e-03
Equipment Usage	3.40225360e-03
Material Quantities	3.26026038e-02
Project Duration (days)	-3.71451824e-02
Resource Allocation Efficiency	-5.75301286e-02
Schedule Optimization	-2.89036644e-02
Computation Time (CT)	-6.12703459e-02
Best Cost (BC)	8.99490743e-02
Evaluation Metric (Nfe)	-2.51589332e-02
Mean Resource Demand	1.83111532e-03
SD of Resource Demand	1.76418385e-02
Risk Level	1.00564596e-01
Intercept	-0.08157324

The results of the logistic regression used to predict the level of risk in construction projects are the coefficients of the model and the intercept as presented in Table 3. The coefficients represent the magnitude and the course of the connection among the characteristics and the possibility of occurrence of the outcome (level of risk). Indicatively, the coefficient of Labor Requirements takes the value of -2.18965383e-03 implying a negative correlation between it and the level of risk meaning that, as labor requirements rise, the level of risk falls to a very small degree. Material Quantities on the other hand, is positive with a coefficient of 3.40225360e-03 which means that the greater the quantities of materials, the more risk of poor performance by project or defaulting of the contractor. The Intercept of -0.08157324 indicates the value of the log-odds of a project at the high-risk level when the predictors are all zero. The coefficients play a key role in predicting the risk and the performance of the project since they measure the effect of various variables of the project.

Table 4 Logistic Regression Model Classification Report

Metric	**0**	**1**	**2**	**Accuracy**
Precision	0.36	0.43	0.32	0.36
Recall	0.16	0.41	0.53	
F1-Score	0.22	0.42	0.40	
Support	102	106	92	300
Macro Average	0.37	0.37	0.34	0.36
Weighted Average	0.37	0.36	0.34	0.36

Table 4 displays the classification report of the logistic regression model to predict the degree of project risk (0, 1, and 2). The model had an accuracy of 0.36, which means that the model had the correct prediction of the risk level 36 percent of the time on all the samples. In class 0 (low risk), the model had a precision of 0.36, recall of 0.16 and a f1 area of 0.22 which means that it has less capability to predict low-risk projects. In class 1 (medium risk), the precision was 0.43, the recall was 0.41 and f1-score was 0.42, which showed slightly higher. In class 2 (high risk), precision was 0.32, recall was 0.53 and f1-score was 0.40 indicating that the model performed better to outline high risk projects. The overall moderate performance of the model, measured in macro average and weighted average at 0.37, value of precision and recall respectively.

Table 5 Confusion Matrix for Logistic Regression Model

Actual\Predicted	**0**	**1**	**2**
0	16	29	57
1	14	43	49
2	14	29	49

The confusion matrix of the project risk level prediction model using the logistic regression model is represented in table 5. The table is presented by the real classes in the rows (0, 1, 2), and the Gerner classes in the columns. In the case of class 0 (low risk), the model forecasted 16 true positives, 29 false positives, and 57 false negatives. In the case of class 1 (middle risk), there were 14 true positives, 43 false positives and 49 false negatives. In the case of the high risk (class 2), the model estimated 14 true positives, 29 false positives, and 49 false negatives. According to the matrix the model is not strong in terms of correct classification in all classes.

Figure 2 below shows the values of feature importance on the logistic regression model, which represents the relative contribution of each feature to their prediction of risk in the project. The highest feature is that of Mean Resource Demand and the strength of this feature was 0.075 which means that this factor has a strong impact on the model predicting the level of risk. Material Quantities and Evaluation Metric (Nfe) also exhibit significant values of 0.025 and 0.05 representing 0.025 and 0.05 respectively. Conversely, such characteristics as Labor Requirements and SD of Resource Demand have little effects on the model with negative values being very close to zero which implies; the less influence they have regarding the prediction of the project risk.

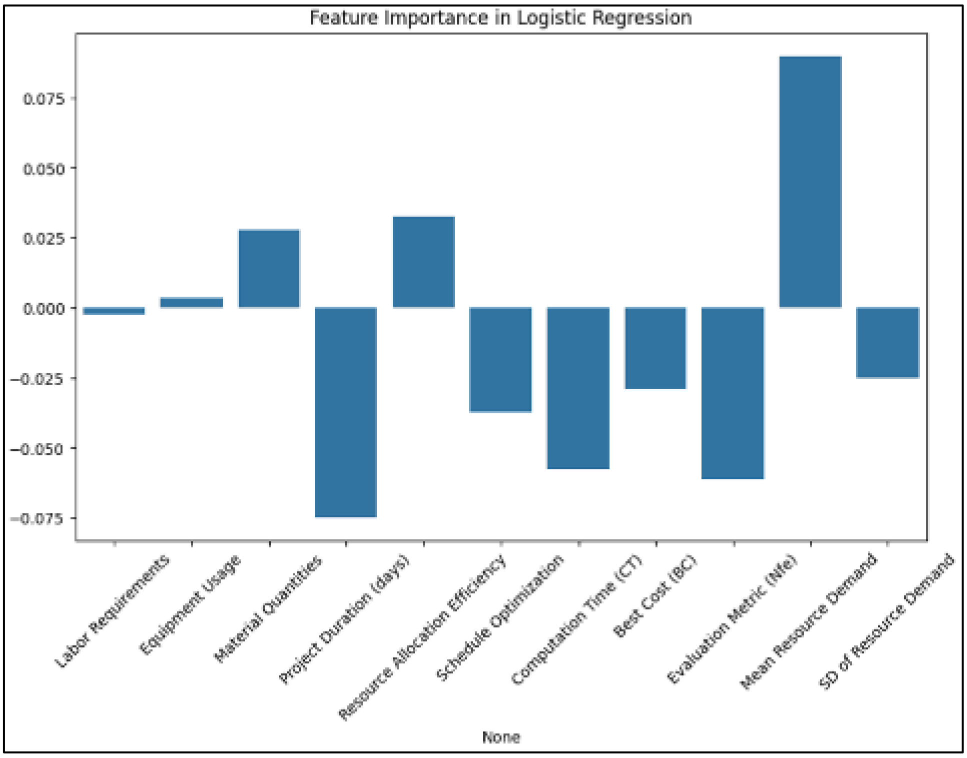

Figure 2 Feature Importance in Logistic Regression

4.1. Implications of Results

The implications of the findings of this study on construction project management are very high. The analysis has shown that project risk prediction is based on such features as Mean Resource Demand and Material Quantities. Through these high-impact factors, those which are managed by the project managers can be addressed proactively to allocate resources and prevent possible delays or cost overruns. Recognizing the level of risks at an early stage during predictive modeling allows using this information to take action and react prior to the risk aggravating challenges and hindering contingency planning. It is an understanding that can be used to better utilize resources, schedule time, and eventually minimize chances of a contractor default or poor performance and achieve more success by delivering projects.

4.2. Limitations

This study has a number of limitations. First, the logistic regression model assumes a linear nature in between the predictors and outcome which may not be able to capture more complex weaker relationships that exist in construction projects. Moreover, the data is secondary and may not represent all the real-life aspects, including external economic or unexpected delays. Moreover, the scope of the data that was used as only a specific group of variables can also limit the generalizability of the model to other construction situations or larger datasets, thus limiting its use to other projects.

5. Conclusion

The purpose of this study was to come up with a predictive risk score model in construction projects based on logistic regression to predict contractor risks and project performance risks. The most important conclusions include the idea that such variables as Mean Resource Demand and Material Quantities are highly predictive of the risk of the project. In spite of certain disadvantages, the logistic regression model revealed that it has effective predicative possibilities to measure the project outcomes using resources allocation and performance measures. The current study is one of the contributions to the expanding range of risk prediction studies in the construction management, and it provides a data-driven method to project managers. To improve on this model in the future, more diverse datasets and more sophisticated machine learning methods such as random forests or neural networks would be beneficial so that there is further refinement to the prediction accuracy and the model strength.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Akinboboye, I., Okoli, I., Frempong, D., Afrihyia, E., Omolayo, O., Appoh, M., Umana, A. U., & Umar, M. O. (2022). Applying predictive analytics in project planning to improve task estimation, resource allocation, and delivery accuracy. International Journal of Multidisciplinary Research and Growth Evaluation, 3(4), 675–689. https://doi.org/10.54660/.ijmrge.2022.3.4.675-689

[2] Assaad, R., El-Adaway, I. H., & Abotaleb, I. S. (2020). Predicting Project Performance in the Construction Industry. Journal of Construction Engineering and Management, 146(5), 04020030. https://doi.org/10.1061/(asce)co.1943-7862.0001797

[3] Bahamid, R. A., Doh, S. I., & Al-Sharaf, M. A. (2019). Risk factors affecting the construction projects in the developing countries. IOP Conference Series: Earth and Environmental Science, 244, 012040. https://doi.org/10.1088/1755-1315/244/1/012040

[4] Bowers, A. J., & Zhou, X. (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. Journal of Education for Students Placed at Risk (JESPAR), 24(1), 20–46. https://doi.org/10.1080/10824669.2018.1523734

[5] Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building Operational data. Frontiers in Energy Research, 9. https://doi.org/10.3389/fenrg.2021.652801

[6] Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. Secondary Analysis of Electronic Health Records, 185–203. https://doi.org/10.1007/978-3-319-43742-2_15

[7] Python Developer. (2025). Construction Project Resource Dataset. Kaggle.com. https://www.kaggle.com/datasets/programmer3/construction-project-resource-dataset.

[8] Rahaman, M. A., Rozony, F. Z., Alam, S., & Haque, Md. N. (2024). BIG DATA-DRIVEN DECISION MAKING IN PROJECT MANAGEMENT: A COMPARATIVE ANALYSIS. Academic Journal on Science, Technology, Engineering & Mathematics Education, 4(3), 44–62. https://doi.org/10.69593/ajsteme.v4i03.88

[9] Raja, K., & Murali, D. (2020). Resource management in construction project. International Journal of Scientific and Research Publications, 10(05), 252–259. https://doi.org/10.29322/ijsrp.10.05.2020.p10130

[10] Shah, F. H., Bhatti, O. S., & Ahmed, S. (2023). Project Management Practices in Construction Projects and Their Roles in Achieving Sustainability—A Comprehensive Review. Engineering Proceedings, 44(1), 2. mdpi. https://doi.org/10.3390/engproc2023044002