www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

APPLYING MACHINE LEARNING FOR PREDICTIVE ANALYSIS IN PROJECT-BASED DATA: INSIGHTS INTO VARIATION ORDERS

SUBMITTED: October 2024
REVISED: May 2025
PUBLISHED: May 2025
EDITOR: Robert Amor
DOI: 10.36680/j.itcon.2025.033

Mirza Muntasir Nishat, PhD Candidate
Norwegian University of Science and Technology (NTNU), Trondheim
mirza.m.nishat@ntnu.no

Aneeq Ahsan, MSc
Norwegian University of Science and Technology (NTNU), Trondheim
Aneeq.Ahsan@dnv.com

Nils O.E. Olsson, Professor
Norwegian University of Science and Technology (NTNU), Trondheim
nils.olsson@ntnu.no

SUMMARY: *The complexity of the global supply chain and project execution necessitates advanced methodologies in project management. As industries are generating large amounts of project data, machine learning (ML) algorithms can be a viable tool for addressing predictive analytics and transforming this industry into more digitalization. This study examines the feasibility of leveraging ML models for predicting variation orders (VOs) in an energy construction project through the use of actual project management data. Using historical project data, this study presents the investigative analysis of applying six ML regression models to predict VOs and evaluates the performance of these models using the mean squared error metric. It is observed that various project activities are nonlinear in the impact of the order of variation, which indicates that advanced ML techniques are required when analyzing the order of variation rather than using linear model analysis. Thus, the results underscore the critical role of ML predictive model implementation in improving change management by enabling preemptive detection of potential problems, risk reduction, and more efficient project execution. Moreover, this study will also help to narrow the existing gap between ML-based theoretical applications and practical project management strategies while also demonstrating the efficacy of AI-based decision support systems for on-time project control. The contributions of this study provide a foundation for developing integrated ML models and project management software, fostering data-driven decision making in dynamic project scenarios.*

KEYWORDS: *Machine Learning, Artificial Intelligence, Project-based data. Project management, Variation Orders, Changes.*

REFERENCE: *Mirza Muntasir Nishat, Aneeq Ahsan, Nils O.E. Olsson (2025). Applying Machine Learning for Predictive Analysis in Project-Based Data: Insights into Variation Orders. Journal of Information Technology in Construction (ITcon), Vol. 30, pg. 807-825, DOI: 10.36680/j.itcon.2025.033*

COPYRIGHT: *© 2025 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

1. INTRODUCTION

Global supply chains and procurement alliances are a driving factor for the contemporary complexity of projects; however, undertakings will become even more challenging to fulfil on time and on budget as civilization continues to grow (Nady et al., 2022). As global economies become increasingly interconnected, project management also has to navigate a complex landscape that needs collaboration at the level of many players. This has created the ever-increasing need to be sustainable, comfortable with risk, and compliant with regulations, driving project complexity and making business as usual ineffective (Turner, 2016). With the increase in the number of projects and the need for performing execution, the need for project management and change management is now in every organization. Other than the financial institution, the overall project environment involves many bilateral agreements and stakeholders; hence, managing a project has become a more and more global approach (Keshavarzian & Silvius, 2022). Within such a framework, industrial strategies should be coordinated for cost optimization and performance maximization (Donyavi et al., 2024). While this presents its challenges, embracing new technologies is an emerging, viable option for organizations facing these complexities.

Technology is advancing towards improving project management efficiency. The popular terms in monitoring and controlling assets are digital twin and multi-agent systems (MAS) (Bahrpeyma & Reichelt, 2022). Similarly, cloud-based collaboration platforms and integrated software solutions are also being combined for optimizing a range of different projects (Bui et al., 2020). Though a huge amount of project data has been collected by industries over the years but project management has not been able to explore the pattern of this data for the implementation of digitalization. This data remains unexplored, although it could be utilized for planning, control, and administration. Currently, most existing project management tools are focused mainly on reporting performance history and on-time performance, but do not perform predictive analysis regarding project variances (Rezvani & Khosravi, 2023). The classical techniques of project estimation find it difficult to accommodate the nonlinearity and complexity that is inherent in modern projects (San Cristóbal et al., 2019). However, these models are generally expensive and time-consuming to build and pose new challenges in predicting project variations with a high temporal resolution, while timelines in projects are inherently porous (Taghaddos et al., 2024).

Despite the movement towards more data being available, the integration of artificial intelligence AI and machine learning (ML) into project management processes is still in its infancy. Many prediction tasks have become impossible, and in some cases, machine learning systems have even begun to outperform human judgment, especially when vast amounts of data and complex patterns are present. For instance, recent studies highlight that artificial intelligence models such as the deep learning based-cost estimation process detect cost overruns, time lags, and schedule deviations better than traditional approaches (Al-Hajj & Ismail, 2018). Such a chasm between data collected vs its application on strategic decisions provides a prospect to optimize traditional project management approaches with ML-backed predictive modelling. Although ML for predictive analytics (Li et al., 2021) has been successfully applied in industries such as finance and healthcare, the construction and engineering industries lag in the deployment of ML for the proactive management of a project.

The current study investigates the applicability of previously identified ML techniques to develop a machine-learning-based model to evaluate the VOs considering project plan data. As such, variation orders are an essential aspect of change management since they influence timelines, funding, and the allocation of resources. VOs are among the main unpredictable challenges in construction and infrastructure works, resulting in low productivity, project delays, and cost overruns (Gündüz et al., 2020). Gaining an understanding of the predictive capability of use cases in the context of ML models would provide the insights needed to optimize project execution. This issue can be posed as a method for predicting how many VOs will take place subsequent week, and flag for the presence of any VOs. By rigorously studying real-world applications, this work contributes to bridging the gap between ML and its use case in project change management. Traditional rule-based systems typically apply generic procedures to assess attack risks and make decisions on responding (Jin et al., 2022). However, ML models can identify hidden patterns and correlations that humans fail to recognize when making decisions.

The main purpose of the pilot study is to investigate the applicability of ML models in predicting project potential variation orders and their impact on the management of projects. Moreover, this investigative approach will create the possibility of exploring the quality of data generated out of any project management software for training ML models afterwards. Thus, the study focuses on the following key factors:

CC BY

- Investigating project data and applying ML algorithms in predicting variation orders (VOs) for the upcoming week.
- Understanding the implications of ML-driven change management and discussing the potential of it for project-specific data.

In the next sections of the study, a background study is portrayed, followed by a description of the methodology for the implementation of ML models. The results are subsequently reported and analyzed with a key findings discussion. Finally, concluding remarks are provided on the implications of this study and possible future research possibilities. Hence, this study conducts a comprehensive investigation of the aforementioned topics to bridge the gap between project data utilization and effective project requirement management through the insights derived using ML.

2. BACKGROUND STUDY

The trend of Artificial Intelligence (AI) applied in project management has attracted considerable attention over the past few years, propelled by the growing demand for data-driven decisions and predictive analytics (Ali et al., 2023). With industries developing extensive project data, AI provides us the capability to utilize ML techniques in project planning, monitoring, and control (Odeh, 2023). Although large datasets are available, the uptake of ML in project-based activities is still low compared to other applications in finance, healthcare, and so on. ML is superior to humans at finding hidden patterns and trends in large datasets, however, the inherent uniqueness of projects makes it difficult to build models (Kreuzberger et al., 2023). Since projects are dynamic, they require constant management of multiple stakeholders and changing requirements, which makes integrating AI into them challenging. This is in contrast to repetitive manufacturing processes, where conditions and stakeholders are constant, and requirements rarely overlap.

The ability to accurately forecast resource management optimization values in project-based industries has become quite an uphill task as AI and ML take on more significance. When viewed under the scope of organization-specific project analytics, AI has significantly improved risk identification capability, cost estimation, and schedule adherence (Rezvani & Khosravi, 2023). AI systems assisting planning in a project allow for real-time decision support: the AI will help in a proactive way to respond to emerging risks. Specific factors such as data quality, interpretable AI models, and domain knowledge must be considered in the successful implementation of AI in business. Adoption of AI has been emphasized in literature as a driver for benefits like automating mundane project work and enhancing decision support systems (Bui et al., 2020).

Several research works have put machine learning models for project management in place across a variety of domains. For instance, Mahdi et al. (2021) use machine learning in the analysis of software success or failure factors. Conversely, Ma et al. (2021) studied the applications of machine learning in locating the risk factors for construction projects and further developments in risk assessment models. On the other hand, Aldana et al. (2021) have shown preemptive approaches to delivering projects based on AI forecasts for a shorter time for completion. Peña et al. (2019) have reviewed project control mechanisms and examined the possibility of using AI in machine learning to find inefficiencies and optimize workflows. In addition, Van et al. demonstrated that ML could be critical to the analysis of project documentation and generating value from textual resources like reports of meeting minutes and project documentation (Van Niekerk et al., 2022). This research was able to indicate an 80% success rate in predicting the completion times of projects using Natural Language Processing (NLP) techniques over archived project data.

Several works have thus managed to employ ML models within project management and across domains. As other examples, Mahdi et al. (2021) talked about embedding machine-learning techniques into software success/failure factors analysis. Ma et al. (2021) investigated machine learning applications in risk identification for construction project factors with a focus on enhancing risk assessment models. On the other hand, Aldana et al. (2021) discussed AI-based project delivery methods for predictive analytics to have a shorter completion time. Peña et al. (2019) presented the use of AI for the project control mechanism analysis, which thus showed that machine learning has great promise in identifying inefficiencies as well as optimizing workflows. Furthermore, Van et. al presented. ML can be involved in documentation analysis of projects as well as driving added value from texts like meeting minute reports and project documentation (Van Niekerk et al., 2022). This research achieved an 80% success rate for completion duration prediction using natural language processing techniques over archival project data.

CC BY

Both project theory and practice highlight the importance of change management, and multiple studies establish the relationship between change management and the success of projects in terms of cost, schedule, and stakeholder satisfaction. Research conducted in the early stages by Kartam (1996) indicated that the identification of changes in projects at early stages was crucial to mitigate conflict and avoid cost escalation. Subsequent work has built on this work, adding machine learning predictive models to project variance orders and their impacts (Gündüz et al., 2020). Zhao et al. (2008) focused on change management frameworks, and this was only mentioned by Lee et al. Potential impacts of changes in projects: (2006) developed dynamic planning models to evaluate the impact of project modifications. Similarly, Assaf and Al-Hejji (2006) showed that variation orders are among the most common factors that impact the project schedules, with many construction projects suffering time overrun as well as cost overrun due to variation orders. It demonstrates the need for integrating predictive modeling techniques to help with unplanned changes on projects.

Although there is an increasing body of literature investigating AI applications in project management, several research gaps still exist. Most studies are concentrated on performance indicators and on time-frame estimation, while the aspect of AI-based change management is less investigated (Himeur et al., 2023). Though industries like software development have adopted the usage of AI-driven methods for predictive analytics, the construction and infrastructure project domains continue to rely extensively on traditional estimation techniques (Li et al., 2021). In addition, prior studies have predominantly investigated the use of AI embedded in large-scale applications, contributing to the limited understanding of how AI affects smaller, domain-specific projects. Very few studies have been conducted on the predictability of AI models developed in variation order analysis, a signaling point towards such studies as identifying, preventing, and limiting changes will lead towards a more analytical view and scope.

Systematic Approach to rigorous evaluation of AI applications in project management key criteria in deriving meaningful insights include ensuring the quality of data, selecting usable ML models, and validating predictive accuracy (Kerzner et al., 2022). Also, AI could have an impact where ethical aspects come into play, since AI-supported decisions may alter the path and results of a project or projects. Cui and Olsson (2009) reported that cost reduction and the improvement of project success rates are the result of well-structured scope change management processes. Their findings also depict that data-driven strategies are critical in mitigating risks and improving the decision-making process. By leveraging Machine Learning (ML) models as a bridge between data-driven forecasting and proactive project control through the integration of these models within the change management process, this study focuses on the following research gaps:

- While there is no shortage of data generated in projects, the introduction of AI and ML into the project management processes is still in its infancy, especially with the use of predictive analytics in change management. The existing literature largely emphasizes project performance metrics over proactive risk mitigation.
- Previous research studies mostly reviewed trends in change management on a project level, while, at the same time, there are very few studies that focus on using a single project to predict the change in orders. This gap in the literature is addressed by this study, since it investigates how ML can be used to enhance the prediction of changes at the project level, enabling improved decision-making and project control.

Thus, the background study lays the groundwork for the implementation of ML models to project variation order analysis. Acting as a magnitude report, determining the research gaps already done and how important predictive modeling is in project management, this paper aims to contribute to the advancement of AI-driven project management methodologies. For future works, researchers can explore improvements in AI technologies, embedding common-sense knowledge, and their relevance in real-life project management applications. Incorporating these insights into project management systems will create models for data-driven decision-making, making AI practical beyond theory.

3. METHODOLOGY

The study is based on data gathered from one large construction company in Norway. The data came from a single project that was stored in project management (PM) software and had sufficient data that could be extracted by executing SQL queries through the software interface. The information in the spatiotemporal (ST) type data was about the level of activity and milestones, where the quantity of recorded variables is a spatial indicative and the

CC BY

project's spontaneous progress is in the time domain (Wang et al, 2020). The paper is based on a larger case study report, which includes further details on the work (Ahsan, 2022).

3.1 Original dataset

Period status tables for resources and activities, a change register, intermittent tables for connecting data, and higher-level data aggregation are all included in the data. While period status tables include historical data on activities with specified attributes being recorded, activity tables transmit a picture of activities at one specific moment in time. In a similar vein, the resources table includes a summary of activities at a given point in time, but the corresponding resource table's period status contains historical data on resources. All the requests for modifications made during the project are included in the change register table, which is linked by the primary and secondary keys to the resources table, which is related to the activities table. Different tables hold a variety of data kinds, including textual information on projects, including the discipline of activity, as well as information about constraints, flags, and computed characteristics.

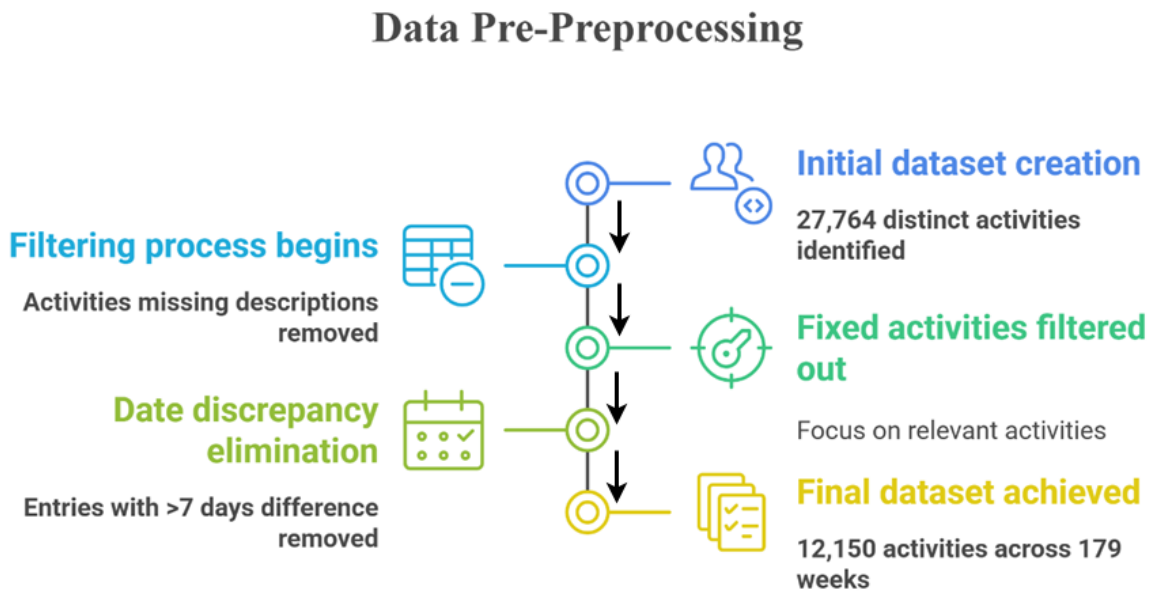

Figure 1: Data pre-processing.

3.2 Data Preprocessing

To ensure that the data quality is appropriate, data preparation is a crucial stage (Brownlee et al, 2020). Reading data that has been taken out of the PM software database and combining it into a single dataset so that it may be processed later is known as data preparation. The activities table's period status was first read into a pandas data frame, paying close attention to the date columns and structure. The activities table was read in the second phase to make sure it had crucial details about the activities, like flags, reference fields, original plan dates, outline codes, etc. Using a left join, the period status data frame and the activities table are combined such that all the information from the period status is kept and just the matching information from the activities table is retrieved. The common primary key for both data frames was used for this join operation.

In the data frame, there were 27764 distinct activities. For instance, the activities table's description column is filtered such that it has a value since the PM program requires a description for each activity in a project. A filter was applied to the dataset so that only those activities that are not fixed activities remain. The period start and cut-off dates are the two crucial date columns in the period status tables that are essential for this investigation. The cut-off date indicates how long it will take before an update on the status is conducted, and the period beginning date indicates when it has been executed. Given that weekly data was the primary emphasis, every entry in the dataset with a difference of more than seven days is eliminated. Following the application of the filters, 12150

CC BY

activities in total and 179 weeks of data were left. According to Corrales et al., if an activity in the PM program is not linked to a calendar, it is viewed as poor planning or activity reporting (Corrales et al, 2018). As a result, these types of activities, which are termed as noise, were eliminated. The data pre-processing steps are illustrated in Figure 1.

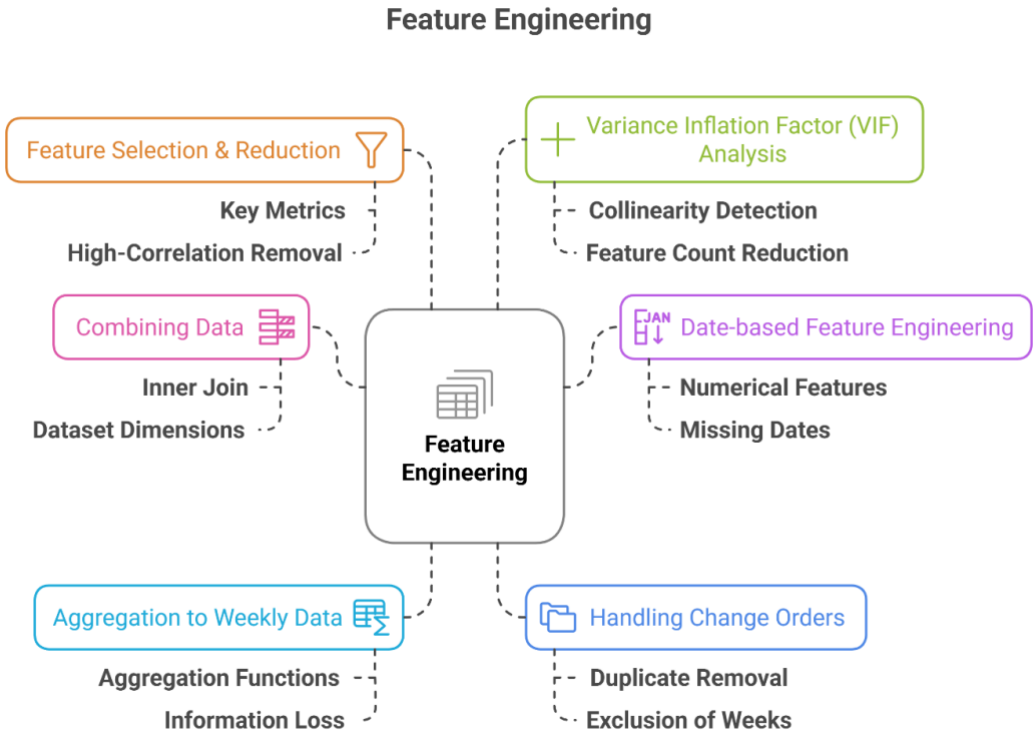

Figure 2: Feature Engineering.

3.3 Feature Engineering

A table including values of interest, such as period earned value, expenditure amount, projected quantity, and contractual volume, represents the period status of resources. After filtering, this data is combined with the resultant dataset using an inner join to keep just the weekly data connected and all the filtered activities from the period status of the activities table. Since the target variable will only be retrieved for actions that are genuinely present in the dataset, this filtered data frame is kept as a reference for extracting the target variable. This data frame can be referred to throughout the study as "filtered_df", and the dimension is '188731 x 99'. Since most machine learning algorithms do not take date-time formats as input for their models, different features for date fields were computed. Features like days till frontline date, days till early start, days till early finish, late finish, current early start, and current early finish are examples of features. To allow the machine learning algorithm to distinguish between missing dates during training, the resulting column was assigned a value of -100 when a date was absent from the data frame. So, the shape of the dataset becomes '188731 x 122'. Because there are varying numbers of activities every week, the data frame is aggregated based on weekly data. Four aggregating functions are used: minimum, maximum, mean, and sum. These consolidation routines take the input data, aggregate it into weekly data, and then apply the designated function to the data. As a result, the data frame's dimensions are now '179 x 529'. However, since regression implies a fixed degree of aggregation (Berry et al, 1993), some information is lost when activity level data is aggregated to weekly data, since each week's aggregation has a variable granularity level.

CC BY

When a change order for a particular activity appears more than once, the duplicates are eliminated. To get only pertinent change orders, the stored activity sequences are retrieved and merged with this data frame using an inner join. The number of activities that have change orders on a certain change order issuance date is used to aggregate the data for change orders. The dataset target column has a lot of zeros and a lot of peaks in some weeks since the change order count variable is the target. Project management best practices dictate that modification orders should not be issued too soon after a project is completed, thus, the final 32 weeks of data are not considered the dimension is reduced to 147 x 529. Several features, including baseline planned quantities, estimate at completion, estimate to complete, expended quantity, original plan quantity, remaining work, periodic scheduling factor, schedule variance, to complete performance index, and earned quantity, can be extracted by users of the PM software through its reporting functionality. The pre-processing stage of the data includes eliminating high correlation factors, multiple collinearity, and duplicate columns. Additionally, strings and columns containing date information are removed, which decreases the size to 147 x 324.

The variable inflation factor, as proposed by García et al., is employed to assess the collinearity of a dataset (García et al, 2015). Before using the variance inflation factor (VIF), each column in the dataset, aside from the target variable, has its Pearson correlation assessed. The VIF target column is removed to apply, and VIF values are computed for every feature. The VIF factor has a threshold of 10, and any column with a VIF value greater than 10 is eliminated. After taking these actions, the number of features, which did not include the target variable, went from 324 to 57. The feature engineering process is presented in Figure 2.

3.4 Data Splitting

Data splitting is an important step in terms of applying machine learning as it ensures that models generalize effectively to new, previously unseen data and provides a more accurate assessment of performance (Raschka and Mirjalili, 2019). By using Python, the data is divided so that the last 20% is utilized for testing and the remaining 80% is used for training and validation.

3.5 Data Scaling

Data scaling, commonly referred to as feature scaling, is a machine learning preprocessing step that entails converting the values of various features, or variables, to a common scale. Ensuring that each feature contributes proportionately to the model procedure of training and preventing features with bigger scales from predominating over smaller-scale features are the main objectives. For algorithms, like gradient-based optimization methods included in many machine learning models, that are sensitive to the scale of input characteristics, scaling is particularly essential. Scaling of data helps algorithms converge faster (Singh and Singh, 2020). After data splitting, a standard scaler is utilized to scale the data.

$$X_{Standardized} = \frac{X - \mu}{\sigma}$$

Where, X = Original value, μ = Mean of the feature, σ= Standard deviation

Standard scaler was used over min-max scaler as it maintains the effect of outliers, scales features to zero to enable stable convergence in ML models, and steers clear of compression problems which occur with min-max normalization. With real-world project data containing large differences, standard scaling ensures more generalizable and explainable model performance.

3.6 Applying ML Model

For applying machine learning algorithm, six different ML techniques have been studied and brought into play in this study which are known as long short-term memory (LSTM), Light Gradient Boosting Machine (LGBM), Extra Trees regressor (ETR), Gradient Boosting Regressor (GBR), Ada Boost regressor (ABR) and Cat Boost Regressor. Long short-term memory (LSTM) is a sort of RNN that utilizes a particular hidden unit in a cell to remember information. It also includes a feedback mechanism that sends outputs from the preceding layer to the entire network. The primary capability of LSTM is its capacity to sequentially capture information across time (Song et al, 2024). GBDT can also be used for light GBM. It was designed to address big data difficulties like the curse of dimensionality. It introduced two enhancements over GBDT. The first is the use of gradient-based one-sided sampling, in which any instance with higher error or loss provides more to the knowledge gain criterion for node

CC BY

splitting (Chen et al, 2023). The Extra Tree Regressor (ETR) constructs an ensemble learner using a classical drop-down approach. It employs a decision tree as a base learner. It accepts two hyperparameters: one specifies how many random features will be selected for each tree, and the other specifies the minimum sample size for splitting a node (Gupta et al, 2023). Gradient Boosting Regressor (GBR) is a particular sort of gradient boosting machine. In this strategy, each base learner's training is dependent on previously trained base learners. These GBMs operate in a way that maximizes the correlation between base learners' negative gradient losses (Sibindi et al, 2023). AdaBoost was created by and is now one of the most often used ensemble algorithms. This technique focuses on the training samples with the most loss (Shanmugasundar et al, 2021). Cat Boost Regressor (CBR) is a gradient-boosted decision tree approach that incorporates two innovations: ordered target statistics and ordered boosting (Ibrahim et al, 2020).

3.7 Hyperparameter Tuning

In the process of developing a machine learning model, hyperparameter tuning is an essential stage. The process entails determining the optimal hyperparameter configuration to guarantee that the model performs optimally on the given task and can be effectively generalized to new data. A variety of hyperparameters, including learning rate, maximum model depth, number of leaves, sub-sample, regularization parameters, and minimum child samples, were tested and optimized.

3.8 Performance Metric

Performance metrics are employed to assess a task's correctness or error rate. It compares actual values to predicted values (Botchkarev et al, 2019). In this analysis, we will utilize the mean squared error (MSE) parameter to track model performance. MSE is a commonly used statistic for regression analysis and count data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where n = Number of data points, Y_i = Observed values, $\hat{Y}$ = Predicted values

In this study, Mean Squared Error (MSE) has been selected as the primary measure of performance over other metrics like R^2, MAE, and MAPE as MSE penalizes larger errors due to the squaring of the residuals, which is crucial in our scenario as extreme deviation predictions (e.g., large order variances in project management) can have obscenely negative impacts on projects. By giving priority to larger errors, MSE keeps the model concentrating not only on average performance but also eliminates outliers, which are particularly unwanted in the case of project risk prediction. On the other hand, R^2 is a commonly used goodness-of-fit metric, but it may be confusing when applied in the assessment of predictive performance. R^2 only expresses how effectively variance in the dataset is explained by the model and does not quantify directly the error magnitude. In the case of non-linear, sequential data like that of project progress over time, R^2 can yield high scores even when predictive accuracy is low. However, for MAE (Mean Absolute Error), while it provides an easily comprehensible average error, it treats all errors linearly without giving any extra consideration to large errors. In project management applications, however, large errors (such as a large unexpected variation order) are much more significant than small ones. Therefore, a measure such as MSE that gives greater weight to larger deviations is preferable to provide tighter control over outlier effects. Similarly, MAPE (Mean Absolute Percentage Error), although useful in some business forecasting applications, is extremely susceptible to low actual values, and this will cause instability in percentage-based error calculations. Due to project performance measurements (e.g., weekly percent progress) sometimes approximating small values, MAPE could produce inflated or undefined errors and hence should not be used for this application. In summary, MSE was selected since it better fits the risk-sensitive objectives of the study, penalizes greater prediction errors, and avoids interpretational as well as stability issues of R^2, MAE, and MAPE. This choice offers a more stringent and more appropriately practical measure of model performance given the predictive project management context.The overall workflow diagram of the method is depicted in Figure 3 (a), and the weekly change order prediction process is presented in Figure 3(b).

CC BY

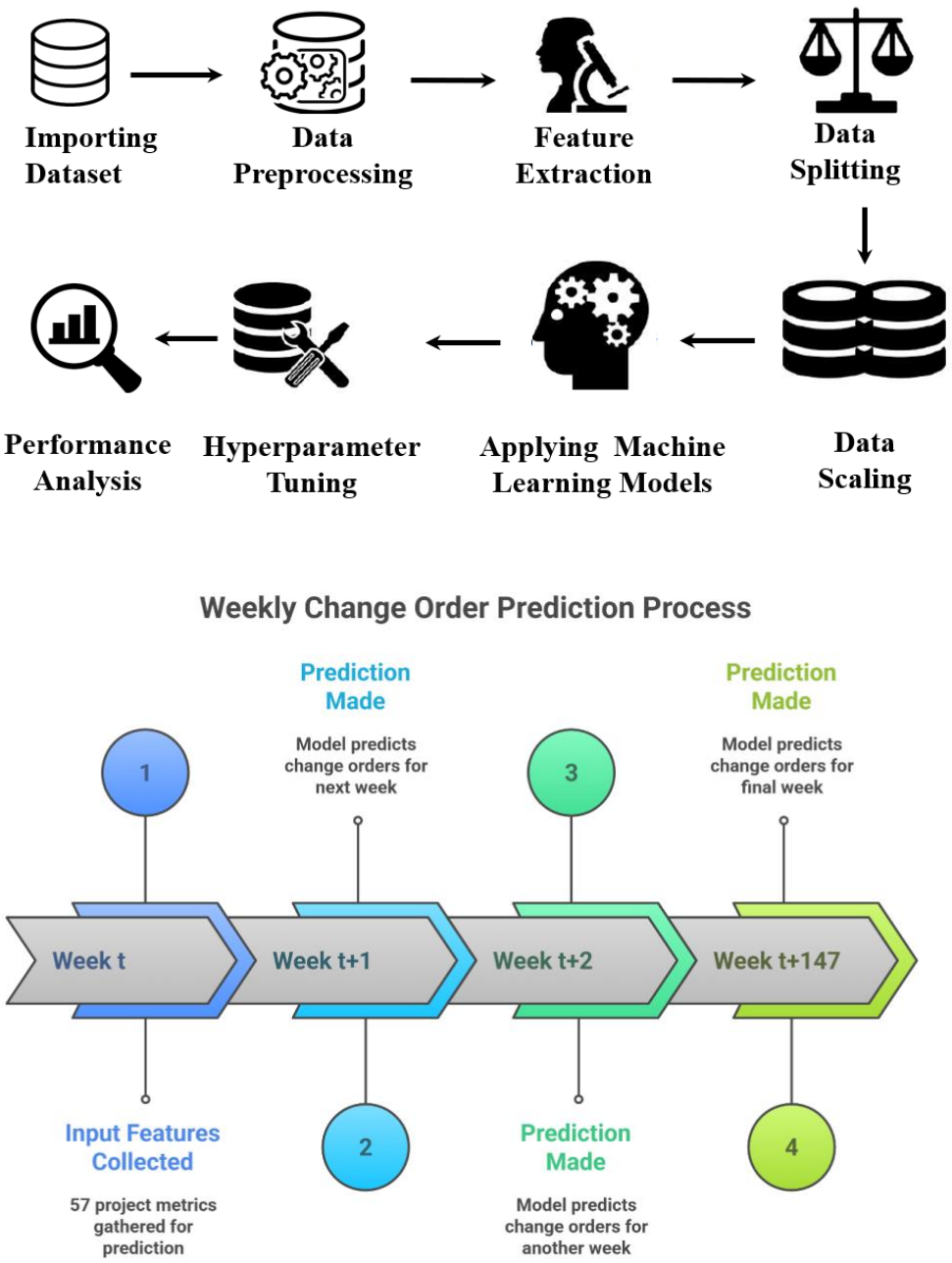

Figure 3: (up) Overall workflow diagram of the proposed methodology (Nishat et al, 2021) (down) Weekly change order prediction process.

4. RESULTS

The purpose of this study is to illustrate a regression analysis using six machine learning algorithms. Table 1 displays the values of Mean Squared Error (MSE) for all the ML models for training and test sets. After performing the simulations, the plots for training and testing data are exhibited. The number of weeks in each set of data is represented by the X-axis, which is termed as index, and the converted number of variation orders' actual and expected values are represented by the Y-axis.

CC BY

Table 1: The values of mean squared error (MSE) for different ML models.

	Data Partition Used	
Model	**Training Set**	**Test Set**
Baseline	1.0224	0.4690
LSTM	0.1351	0.7399
LGBM	0.0371	0.7592
ETR	0.00	0.6143
GBR	0.0933	1.3289
ABR	0.7508	0.4028
CBR	0.0035	0.4003

4.1 LSTM Model

First, a long short-term memory (LSTM) model was applied with hyperparameter optimization (Song et al, 2024). It was realized that a single 400-cell LSTM layer connected to a dense layer of 100 neurons and a final layer of one cell with a linear activation was the optimal arrangement. To compare test results and assess how well the results fit the training data and test data, the findings were shown in Figure 4 and Figure 5, respectively.

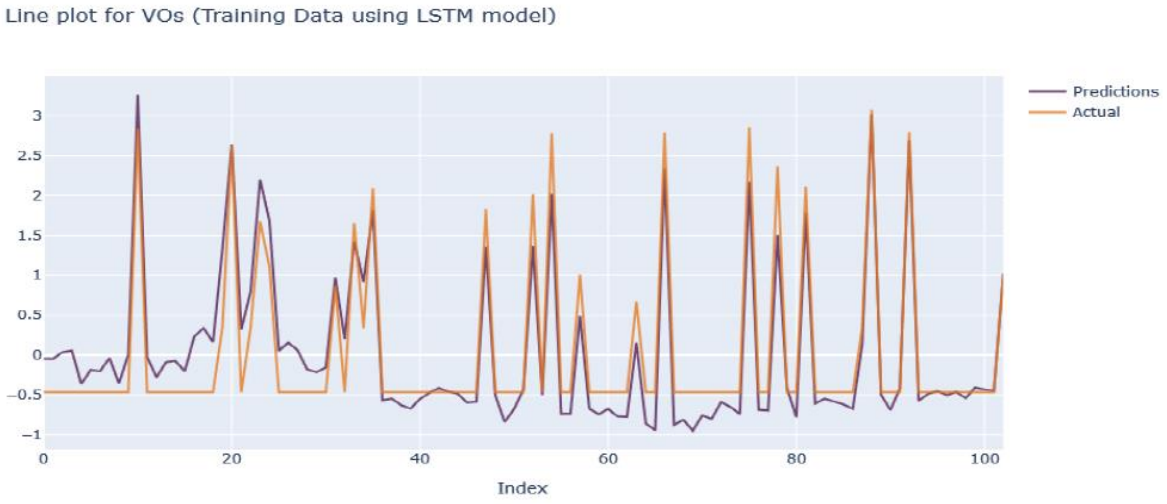

Figure 4: Performance of the LSTM Model using training data.

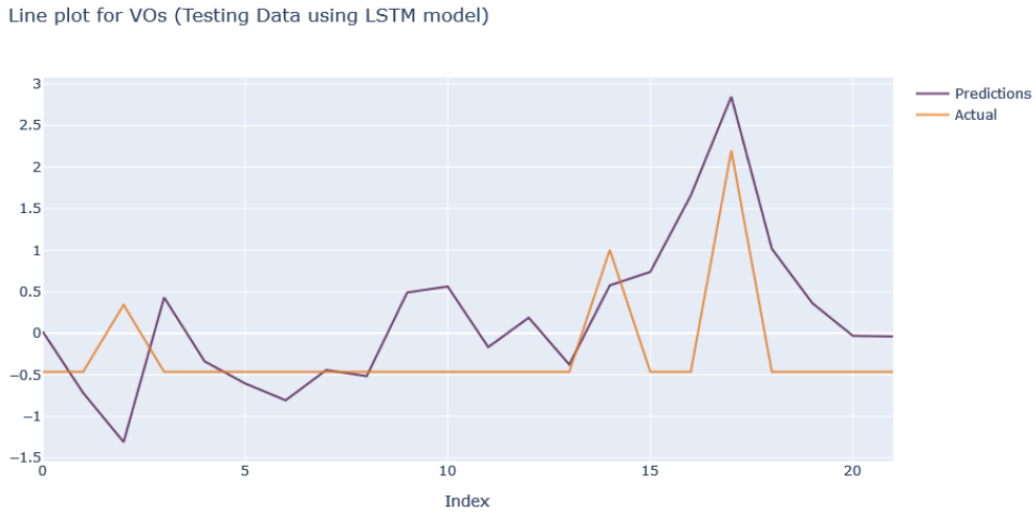

Figure 5: Performance of the LSTM Model using testing data.

4.2 LGBM Model

Second, Light Gradient Boosting Machine (LGBM) was deployed, and it was observed that it did not depend on data scaling, and it had been tailored for hyperparameters utilizing unscaled data (Chen et al, 2023). There were 100 optimization cycles completed, and the training set's lowest loss was recorded after each round. The evaluation rounds' saved parameters were utilized to extract the optimal model. Figures 6 and Figure 7 show a graphic comparison of these outcomes for the training and test sets.

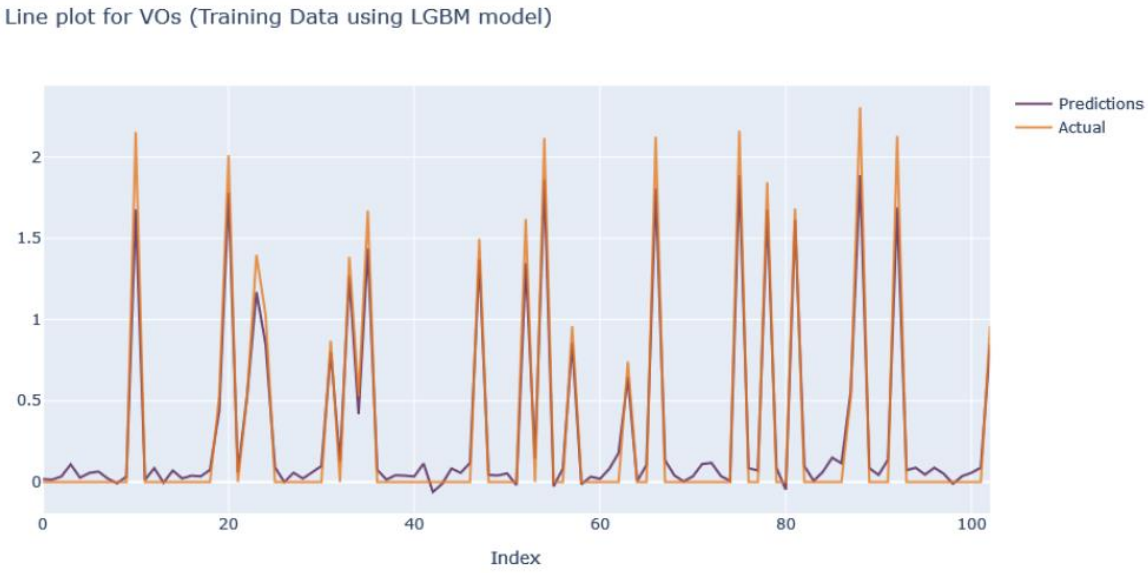

Figure 6: Performance of the LGBM Model using training data.

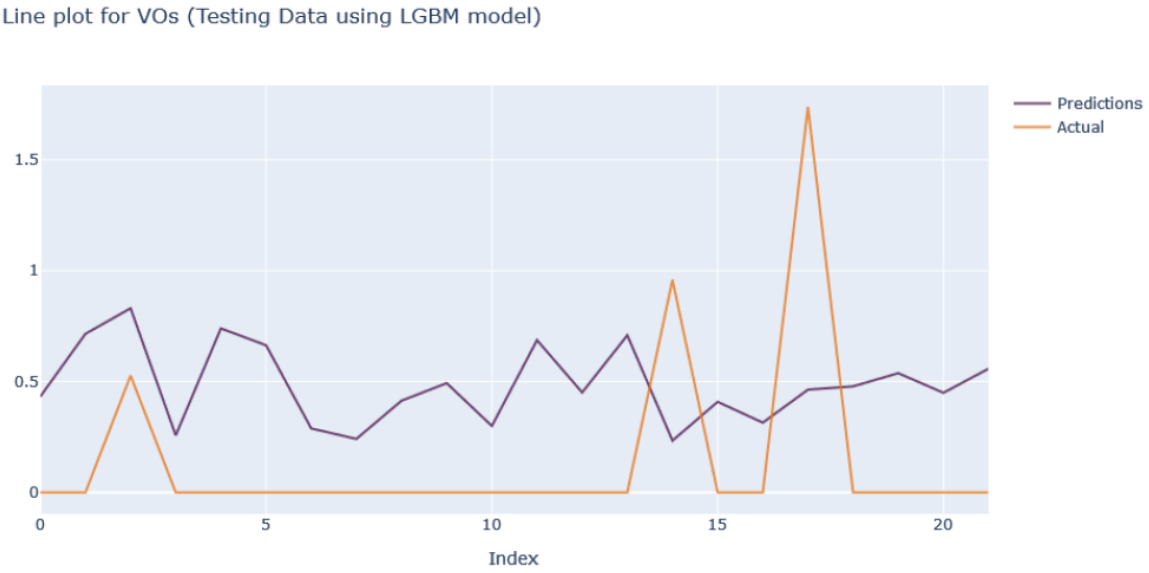

Figure 7: Performance of LGBM Model on testing data.

4.3 ETR Model

Third, the Extra Trees regressor was brought into action and hyperparameters were tuned with 20 iterations to minimize validation error (Gupta et al, 2023). Since the data completely matches the training data, the comparison of the training data is not displayed on a graph. Figure 8 illustrates the plotted graph used to test the data and visually assess the fit.

Figure 8: Performance of the ETR model using testing data.

4.4 GBR Model

Fourth, Gradient Boosting Regressor (GBR) was implemented where the training of each base learner is dependent on other base learners that have already been trained (Sibindi et al, 2023). Figures 9 and 10 display the graphical comparison for training data. And testing data respectively for the GBR model.

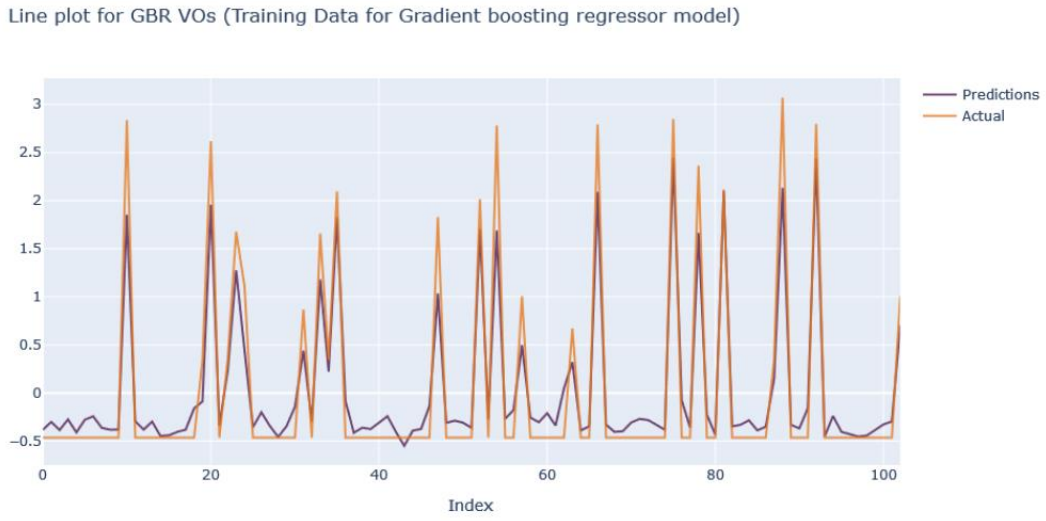

Figure 9: Performance of the GBR model on training data.

4.5 ABR Model

Fifth, Ada Boost regressor (ABR) was developed, which is referred to as one of the most widely employed ensemble techniques (Shanmugasundar et al, 2021). Using validation data, the AdaBoost regressor (ABR) was trained, and the hyperparameters were optimized for the lowest mean squared error. Figures 11 and 12 display the performance of the ABR model for the training and testing sets, respectively.

CC BY

Figure 10: Performance of the GBR model on testing data.

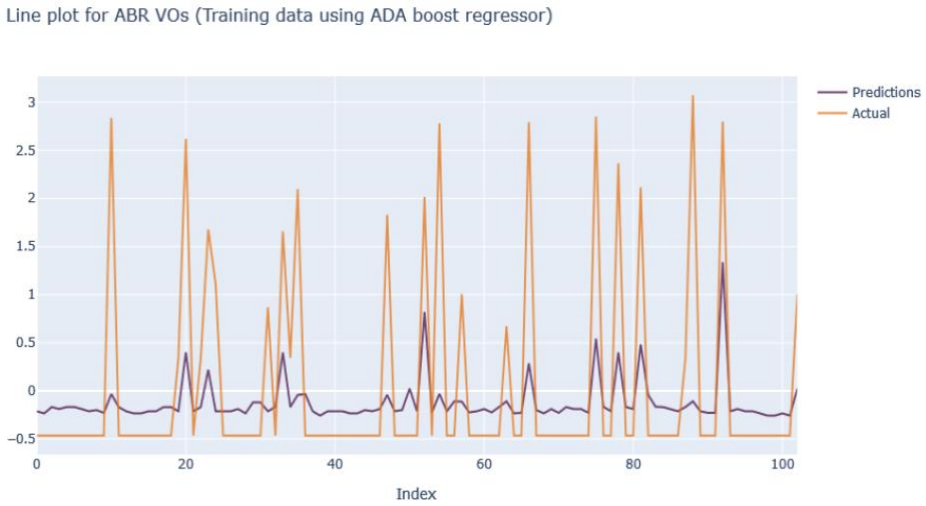

Figure 11: Performance of the ABR model using training data.

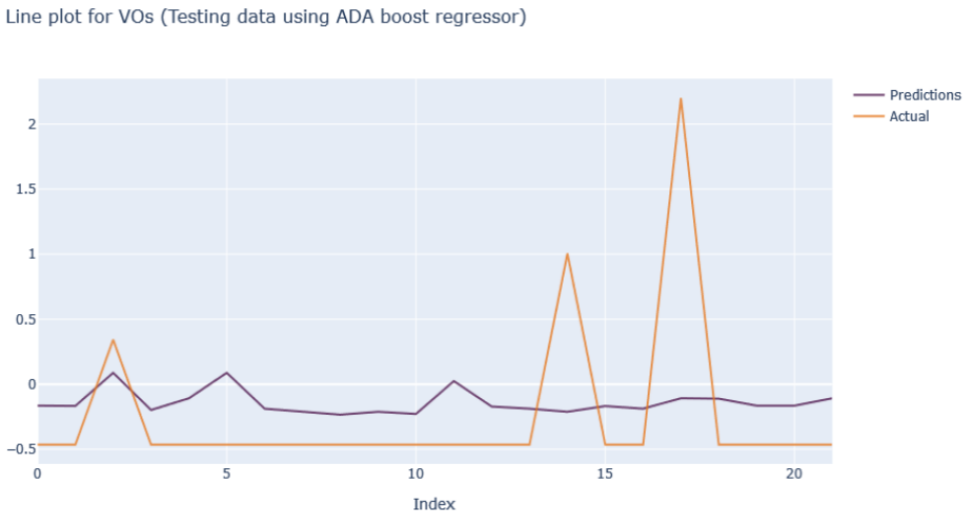

Figure 12: Performance of ABR model using testing data.

4.6 CBR Model

Lastly, the Cat Boost Regressor (CBR) model was carried out by measuring the mean absolute error on the training set and assessing the fit on validation data, and was optimized for training loss (Ibrahim et al, 2020). For hyperparameter optimization, a total of twenty iterations were completed. The testing data was plotted and is displayed in Figure 13 since the training data is almost perfectly matched.

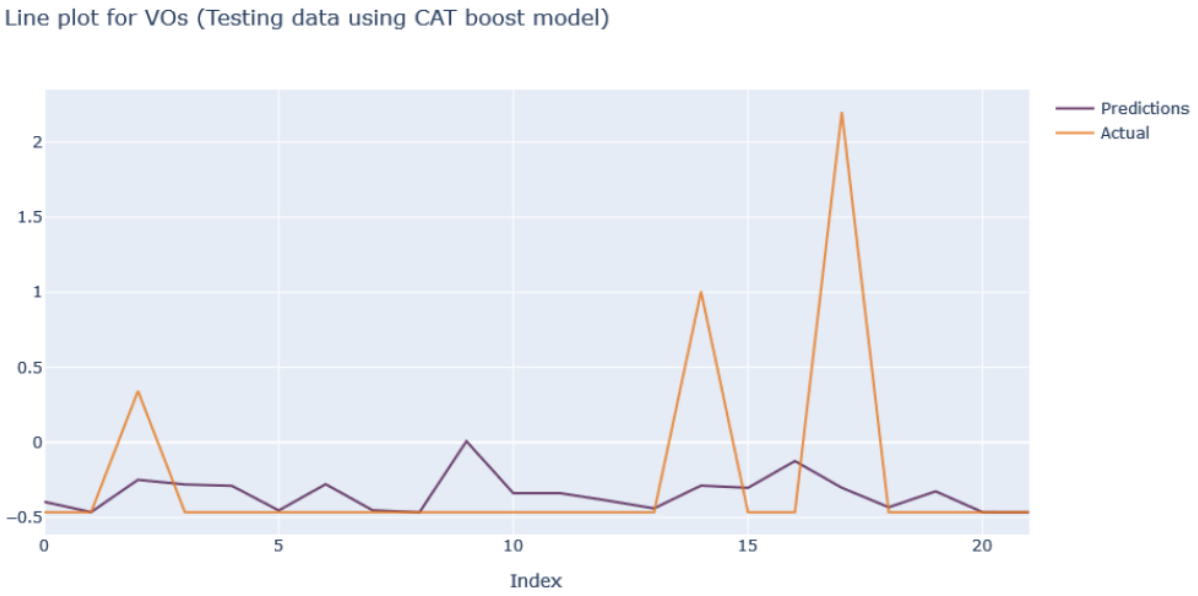

Figure 13: Performance of the CBR model on testing data.

There could be various reasons why a machine learning model's outcome on a test set is often inferior to that on a training set. Overfitting appears to be one of the most common problems, occurring when a model learns the training data too well, including noise and outliers, and then fails to generalize to new, unknown data. As a result, the model may excel on the training set but struggle on the test set. However, insufficient data can also affect regression analysis predictions. If there is a limited amount of data available for training, the model may not understand the underlying patterns well enough to generalize to new samples. Furthermore, hyperparameter adjustment is important for getting optimal model performance. Furthermore, model complexity can capture noise in training data rather than underlying patterns, influencing predictions. Simplifying the model or utilizing regularization techniques can help avoid this. If the metric used to evaluate the model during training is not reflective of the model's performance on the test set, the findings may not be generalizable. To increase performance on the test set, these difficulties must be addressed carefully using approaches such as cross-validation, hyperparameter tweaking, feature engineering, and a thorough understanding of data distribution.

5. DISCUSSION

This study presents a weeklong aggregation of values from activities and resources, and after the implementation of ML models, it is witnessed that LSTM outperformed the other conventional machine learning algorithms (additional trees, LGBM, and gradient boosting techniques). Despite LSTM showing a higher generalization capability (more adaptable to unseen data) than the other categories, the CatBoost model is more suitable for model interpretation. Hoever, previous literature provides evidence that machine learning can be integrated in project management across different dimensions, including but not limited to risk assessment (Ma et al., 2021), failure analysis (Mahdi et al., 2021), project delivery optimization (Aldana et al., 2021), and documentation mining (Van Niekerk et al., 2022); however, the present paper takes a novel approach and provides week-by-week spatiotemporal modeling of change order prediction, using structured data extracted from project management software. Unlike Peña et al. (2019), who concentrate on project inefficiency diagnosis, or Mahdi et al. (2021), who study software success factors, here, the work focuses on the predictive forecast of scope changes based on empirical data. The methodological novelty of the work is in the dynamic aggregation of 57 designed features for

147 weeks and the combination of models like LSTM and CatBoost and temporal variables, such that the method supports both sequential pattern recognition and interpretation. Therefore, the policy recommendation is aimed at enhancing the current state of policy, i.e., the current project context (e.g., construction project environment), to provide project managers with predictive insight into change orders through a scalable, data-driven decision-support method, compared to past reactive or static models existing within the sector.

5.1 Theoretical Contributions

The findings of this study improve the theoretical knowledge of change prediction in the field of project management through machine learning algorithms. Notably, with the discrete and integrated use of several ML techniques, this study indicates the validity of data-driven decision-making with respect to project scope management. This extends previous work by showing that LSTM models effectively capture sequential dependencies in order of variation data. It puts forward a probabilistically predictive framework as a very different alternative to traditional deterministic change management. Furthermore, the study contributes to the theory of its field by demonstrating the importance of the temporal features in predicting project variations, which previous studies paid little attention to.

5.2 Practical implications and contributions

Thus, changes in scope can adversely impact a project with some delays and an increase in cost and time. It is recommended to make planned changes to the project (Madhuri et al., 2018). Project success is often constrained by project managers' poor understanding of the project scope, especially in the early stages of the project, which limits the ability of domain experts to anticipate change. This finding highlights the need to treat ML models as decision-support tools rather than standalone predictive systems, so experts can step in when needed. This paper aims to contribute to the improvement of project risk management through predictive knowledge. ML-based predictions enable project managers to take proactive steps to position risk on variation orders, which serves as the basis for improved cost and scheduling management.

5.3 Limitation

The limitations of the techniques employed are due to the features of the data and the machine learning techniques applied in the analysis. The data limit is one of the challenges that we can see that relates to the study's project. The first data limitation addressed in this analysis comes from the project data used for completing this analysis, as it was sourced from only one organization. That said, you cannot use this approach to all project data. In addition, this data is limited to the energy building sector, hence, transfer functions and modifications are needed to implement these models in a different domain. It demonstrates the problem of model generalizability — ML models trained on a limited dataset may not work well when applied to project environments of all different shapes and kinds. Multi-project datasets should be studied in the future to increase the predictive robustness of the models. Extracting and transforming complex information was not feasible due to time constraints and the complexity of the data structures for various tables, as well as the focus of the study, which was not to exhaustively extract from PM software but, rather, to focus on relevant features. Feature extraction is a general problem in many types of ML-based project management, and future research can use automated feature engineering techniques.

5.4 Directions for Future Research

The ability to anticipate changes in their projects when making plans may be of utmost importance to project managers. This is where change management can help to mitigate the negative effects of scope changes. Integrating ML predictions into project workflows enables organizations to move from reactive to predictive change management, improving efficiency and cost control. If a successful machine learning algorithm for VO prediction were able to be used, then one of the principal effects could be an increase in labor productivity. Good variance management can mitigate the impact of these VOs, as we can anticipate the expected changes in the future by extrapolating the model to predict the cumulative hours of changes. Such prediction ability can aid in resource allocation and workforce planning, which in turn enhances overall project performance. Future work could investigate combining these models into project scheduling software that can update dynamically and provide real-time change risk assessments during execution

6. CONCLUSION

This study investigates the opportunity for ML-based project management, laying the groundwork for solution decisions. This first major finding from the study based on the PM software leads to the conclusion that better and more structured data should be stored in the PM software to facilitate data handling and accommodate ML automation. The current versions of these software packages severely limit the prospects of directly implementing ML. The insufficient standardized data handling approach employed by PM software presents a major challenge, which therefore requires improvement in data structuring for effective ML integration. Due to these complexities of the software packages, time-wise, a great deal is spent on data extraction, loading, pre-processing, and target construction. Contractors using these software packages must also be careful with entering data into the PM software programs to ensure the highest level of data validity and dependability. After the moment of generating data, the integrity of the data production system is critical for machine learning algorithms, for example, to ensure that the resulting data complies with the characteristics and is valid for the dataset.

Moreover, it is observed that variation order prediction (VOP) issues have been a matter of discussion in previous decades among project managers. With an ML model and a structured data process, VO can be predicted with more accuracy; however, the design of the model and data-generating process are crucial to building accurate systems. Such predictive performance illustrates the potential of ML for optimizing project change management and providing a data-driven alternative to conventional forecasting techniques. Additionally, the model places varying emphasis on activities at different stages of a project. At the beginning, it places more emphasis on actions that were nearly completed, but as the project draws to a close, the actions that are least done are most crucial for the VO to be issued. This is a dynamic weighting approach that helps ML models to adapt to the evolving nature of project risks over time.

However, this study provides an insight into the applicability of ML algorithms for change management and refers to the evaluation of the data predictive power of a single project. The pilot study contributes to a better understanding of how varying project progress can influence which amendments are disseminated and implemented during the project. With data and ML techniques, project management creates a transition from reactive change management by taking proactive risk mitigation measures that allow most efficient project implementation. Additionally, it suggests a dynamic modification for change management at various project phases. On the other hand, the VO predictions from a single project plan data incur covariate shift as well as associated challenges in data management. To address this limitation, future work could evaluate hybrid models that integrate transfer learning techniques to deliver reliable predictions for heterogeneous projects.

Furthermore, this study provides empirical evidence that limited data can produce interesting predictions paired with ML frameworks. If there are several projects available to train the model on, the individuals responsible for project planning can make use of these models powered by machine learning to effectively schedule a project and avoid unnecessary changes in the scope and overall uncertainty before the project starts. Using a large, multi-project dataset increases the potential for these predictive models to be operationalized within lived contexts of project management.

By incorporating machine learning into change management, this evidence-based study supports the theory and has practical implications. In theoretical terms, this study contributes to an improved understanding of how LSTM-based models capture temporal dependencies in project variation data, thus providing a novel probabilistic framework for change prediction. From a pragmatic perspective, this study equips project managers with predictive insights that can improve their ability to manage risk, estimate costs, or schedule their projects. Foresight in project management, organizations to improve the efficiency of project outcomes by transitioning from traditional reactive management to data-driven foresight, thus minimizing the limits of uncertainty. Thus, this study utilizes dynamic sequential learning to predict future changes ahead of emergence, in contrast to previous studies, which emphasize static historical analyses. ML-driven forecasting is now combined with real-time project data to transform project management practices. This research offers insight into how predictive analytics can inform project implementation and provides a forward-looking perspective that can influence future industry practices.

REFERENCES

Ahsan, A. (2022). AI in projects-an investigation of use of project-based data for prediction of changes. Master's thesis, NTNU

Aldana, A. et al. (2021). Exploring the use of artificial intelligence (AI) solutions to improve the accuracy of project delivery forecasts. National Academics (No. 018)

Ali, S. et al. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. DOI: 10.1016/j.inffus.2023.101805.

Assaf, S. A. & Al-Hejji, S. (2006). Causes of delay in large construction projects. International Journal of Project Management, 24(4), 349-357. DOI: 10.1016/j.ijproman.2005.11.010

Bahrpeyma, F. and Reichelt, D. (2022). A review of the applications of multi-agent reinforcement learning in smart factories. Frontiers in Robotics and AI, 9, 1027340. DOI: 10.3389/frobt.2022.1027340

Berry, W. D. (1993). Understanding regression assumptions (Vol. 92). Sage

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 045-076. DOI: 10.28945/4184

Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

Chen, H. et al. (2023). Shield attitude prediction based on Bayesian-LGBM machine learning. Information Sciences, 632, 105-129. DOI: 10.1016/j.ins.2023.03.004

Choi, S.-W. et al. (2021). The Engineering Machine-Learning Automation Platform (EMAP): A Big-Data-Driven AI Tool for Contractors. Sustainable Management Solutions for Plant Projects. Sustainability, 13(18), 10384. DOI: 10.3390/su131810384

Corrales, D. C. et al. (2018). How to address the data quality issues in regression models: a guided process for data cleaning. Symmetry, 10(4), 99. DOI: 10.3390/sym10040099

Cui, Y. & Olsson, N. O. (2009). Project flexibility in practice: An empirical study of reduction lists in large governmental projects. International Journal of Project Management, 27(5), 447-455. DOI: 10.1016/j.ijproman.2008.07.007

Donyavi, S. et al. (2024). Understanding the complexity of materials procurement in construction projects to build a conceptual framework influencing supply chain management of MSMEs. International Journal of Construction Management, 24(2), 177-186. DOI: 10.1080/15623599.2023.2267862

Falkner, S. et al. (2018). BOHB: Robust and efficient hyperparameter optimization at scale. International Conference on Machine Learning.

García, C. et al. (2015). Collinearity: revisiting the variance inflation factor in ridge regression. Journal of Applied Statistics, 42(3), 648-661. DOI: 10.1080/02664763.2014.980789

Gupta, R. et al. (2023). A robust regressor model for estimating solar radiation using an ensemble stacking approach based on machine learning. International Journal of Green Energy, 1-21. DOI: 10.1080/15435075.2023.2276152

Himeur, Y. et al. (2023). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. Artificial Intelligence Review, 56(6), 4929-5021. DOI: 10.1007/s10462-022-10286-2

Hsu, M.-W. et al. (2021). Identifying Inter-Project Relationships with Recurrent Neural Networks: Towards an AI Framework of Project Success Prediction. British Academy of Management.

Ibrahim, A. A. et al. (2020). Comparison of the CatBoost classifier with other machine learning methods. International Journal of Advanced Computer Science and Applications, 11(11). DOI: 10.14569/ijacsa.2020.0111190

Kartam, N. A. (1996). Making effective use of construction lessons learned in project life cycle. Journal of Construction Engineering and Management, 122(1), 14-21. DOI: 10.1061/(ASCE)0733-9364(1996)122:1(14)

CC BY

Kerzner, H. (2022). Project Management: A Systems Approach to Planning, Scheduling, and Controlling. John Wiley & Sons

Keshavarzian, S. & Silvius, G. (2022). The perceived relationship between sustainability in project management and project success. The Journal of Modern Project Management, 9(3).

Kreuzberger, D. et al. (2023). Machine learning operations (mlops): Overview, definition, and architecture. IEEE Access. DOI: 10.1109/ACCESS.2023.3262138

Lee, S. H. et al. (2006). Dynamic planning and control methodology for strategic and operational construction project management. Automation in Construction, 15(1), 84-97. DOI: 10.1016/j.autcon.2005.02.008

Ma, G. et al. (2021). Safety risk factors comprehensive analysis for construction project: Combined cascading effect and machine learning approach. Safety Science, 143, 105410. DOI: 10.1016/j.ssci.2021.105410

Madhuri, K. L. et al. (2018). A triangular perception of scope creep influencing the project success. International Journal of Business Information Systems, 27(1), 69-85. DOI: 10.1504/IJBIS.2018.088571

Mahdi, M. N. et al. (2021) "Software project management using machine learning technique—A Review." Applied Sciences, 11(11), 5183. DOI: 10.3390/app11115183

Memon, A. H. et al. (2010). Factors affecting construction cost in Mara large construction project: perspective of project management consultant. International Journal of Sustainable Construction Engineering and Technology, 1(2), 41-54. Available at: https://penerbit.uthm.edu.my/ojs/index.php/IJSCET/article/view/62 (Accessed: 13 September 2024).

Moselhi, O. et al. (2005). Change orders impact on labor productivity. Journal of Construction Engineering and Management, 131(3), 354-359. DOI: 10.1061/(ASCE)0733-9364(2005)131:3(3)

Nady, A. E. et al. (2022). Factors affecting construction project complexity. The Egyptian International Journal of Engineering Sciences and Technology, 37(1), 24-33. DOI: 10.21608/eijest.2021.96807.1100

Nishat, M. M. et al. (2021, December). Performance Assessment of Machine Learning Classifiers in Detecting Psychological Impact of Postgraduate Students due to COVID-19. In 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) (Vol. 6, pp. 1-6). IEEE. DOI: 10.1109/ICRAIE52900.2021.9703997

Odeh, M. (2023). The Role of Artificial Intelligence in Project Management. IEEE Engineering Management Review. DOI: 10.1109/EMR.2023.3309756

Oyewobi, L. O. et al. (2016). Analysis of causes and impact of variation order on educational building projects. Journal of Facilities Management, 14(2), 139-164. DOI: 10.1108/JFM-01-2015-0001

Peña, A. B. et al. (2019). Method for project execution control based on soft computing and machine learning. In 2019 XLV Latin American Computing Conference (CLEI) (pp. 1-7). IEEE. DOI: 10.1109/CLEI47609.2019.235097

Pospieszny, P. et al. (2018). An effective approach for software project effort and duration estimation with machine learning algorithms. Journal of Systems and Software, 137, 184-196. DOI: 10.1016/j.jss.2017.11.066

Raschka, S. & Mirjalili, V. (2019) "Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2." Packt Publishing Ltd.

Shanmugasundar, G. et al. (2021) "A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining." Processes, 9(11), 2015.

Sibindi, R. et al. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5(4), e12599. DOI: 10.1002/eng2.12599

Singh, D. & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524. DOI: 10.1016/j.asoc.2019.105524

Song, Y. et al. (2024). Modelling and forecasting high-frequency data with jumps based on a hybrid nonparametric regression and LSTM model. Expert Systems with Applications, 237, 121527. DOI: 10.1016/j.eswa.2023.121527

San Cristóbal, J. R. et al. (2019) Complexity and project management: Challenges, opportunities, and future research. Complexity, 2019. DOI: 10.1155/2019/6979721

Sawadogo, P. & Darmont, J. (2021) On data lake architectures and metadata management. Journal of Intelligent Information Systems, 56, 97-120. DOI: 10.1007/s10844-020-00608-7

Taghaddos, M. et al. (2024). Optimized variable resource allocation framework for scheduling of fast-track industrial construction projects. Automation in Construction, 158, 105208. DOI: 10.1016/j.autcon.2023.105208

Taylor, P. (2021). AI and the Project Manager: How the Rise of Artificial Intelligence Will Change Your World. Routledge. DOI: 10.4324/9781003175063

Van Niekerk, J. et al. (2022). The value of data from construction project site meeting minutes in predicting project duration. Built Environment Project and Asset Management, 12(5), 738-753. DOI: 10.1108/BEPAM-03-2021-0047

Wang, S. et al. (2020). Deep learning for spatio-temporal data mining: A survey. IEEE Transactions on Knowledge and Data Engineering, 34(8), 3681-3700. DOI: 10.1109/TKDE.2020.3025580

Waring, J. et al. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine, 104, 101822. DOI: 10.1016/j.artmed.2020.101822

Zhao, Z.-Y. et al. (2008). Applying dependency structure matrix and Monte Carlo simulation to predict change in construction project. 2008 International Conference on Machine Learning and Cybernetics. DOI: 10.1109/ICMLC.2008.4620489

CC BY