Timing is Everything: Defining Project Review Periods Through Monte Carlo Simulation and Machine Learning

ENBIS-25 Conference

André Henrique Alves Carneiro | Linda Lee Ho

September 2025

Production Engineering Department
University of São Paulo, SP, Brazil

Content

Introduction

Introduction: Statistical project control

- Field focused on collecting and analyzing project data,
- Indicators are obtained and compared the result to the baseline/criteria,
- Decision: executing timely corrective actions when necessary.

Three main methods:

- Earned Value Management
- Earned Schedule Management
- Earned Duration Management

Earned Value Management (EVM)/ Earned Scheduled Management (ESM):

- The Time Performance Index (TPI) is the main indicator for EVM
- The Schedule Performance Index (*SPI*) is the main indicator for ESM;
- Both cost-based metric;
- Thresholds relied on expert-definition;
- *SPI* converges to one as the project progresses, potentially misleading stakeholders.

Advances with Earned Duration Management (EDM):

- The Duration Performance Index (DPI) is based on schedule.
- Improved time performance analysis
- But it still depends on expert-defined thresholds, which can affect objectivity and consistency.

Introduction: Emerging data-driven approaches

Recent studies have investigated

- **Statistical Control Charts** and
- **Artificial Intelligence (AI)**

as more robust alternative methods for identifying deviations and supporting decision-making in project monitoring.

Research Problem, Objectives

Research Problem

- When should a project be reviewed to enable early delay prediction without compromising predictive performance?

Research Objectives

- Identify critical project **review periods**.
- Determine the most effective **preprocessing techniques** and **machine learning models** for project delay forecasting.

Literature Review

Project Monitoring Methodologies

Project monitoring has significantly evolved over the past decades, starting with Earned Value Management (EVM), advancing through statistical control methods, and now incorporating machine learning and artificial intelligence techniques.

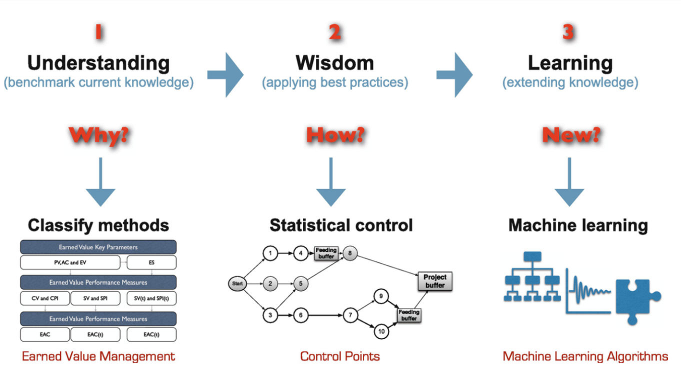

Figure 1: Project monitoring evolution (Source: Vanhoucke (2023))

Earned Value, Schedule and Duration Methodologies

Earned Value Management (EVM), Earned Schedule Management (ESM) and Earned Duration Management (EDM) are all control system, which aims to monitor the project time and cost execution to implement appropriate corrective actions, if necessary.

There is a set of variables and parameters used to calculate their indexes:

- **Planned Value (PV_t)**: is the *cumulative planned cost* for the planned work from the beginning of the project to the review period t according to the baseline schedule.
- **Earned Value (EV_t)**: represents *the accumulated planned cost* of accomplishing the total work performed from the beginning of the project to the review period t.
- **Total Planned Duration (TPD_t)**: is *the cumulative planned time* for all activities until time review period t.
- **Actual Cost (AC_t)**: represents *the value of what has been actually spent* to achieve the progress performed from the beginning of the project to the review period t.
- **Total Earned Duration (TED_t)**: is the sum of the *total time of the activities executed* until the moment of review t.

Earned Duration Methodology

The EDM methodology consists on calculating the Earned Duration (ED_t) and the Duration Performance Index (DPI_t). It follows the same principles of EVM and ESM, but only using time-based metrics:

- $ED_t = t_0 + \dfrac{(TED_t - TPD_{t_0})}{(TPD_{t_{0+1}} - TPD_{t_0})}$: which point of time the project theoretically is.
- $DPI_t = \dfrac{ED_t}{t}$: proportion in which the project is ahead or below schedule, based on duration metrics.

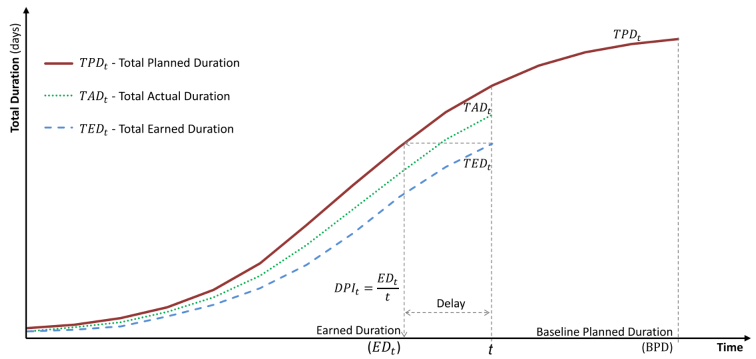

Figure 2: EDM graphical representation (Source: Votto et al. (2020))

Several algorithms have been applied to predict project delays and final costs, including:

- Neural Networks (NN), Support Vector Machines (SVM), and tree-based algorithms
- Nearest Neighbor and anomaly detection techniques
- Preprocessing methods such as Principal Component Analysis (PCA)

Beyond time and cost forecasting, AI has also been used in other areas of project management:

- Resource allocation and scheduling
- Risk prediction and cost estimation
- Communication and decision support

Boruta algorithm has been used a feature selection method, following the nine-step approach (Kursa & Rudnicki 2010)

Methodology

The **Planning phase** focuses on modeling uncertainty, often using Monte Carlo simulation to understand project behavior:

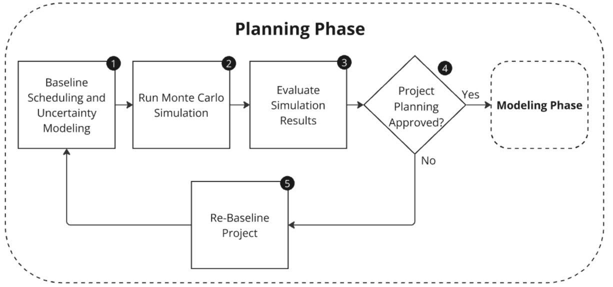

Figure 3: Planning Phase Flowchart.

Modeling Phase

The **Modeling phase** involves applying machine learning techniques to identify key review periods and develop predictive models for project delays:

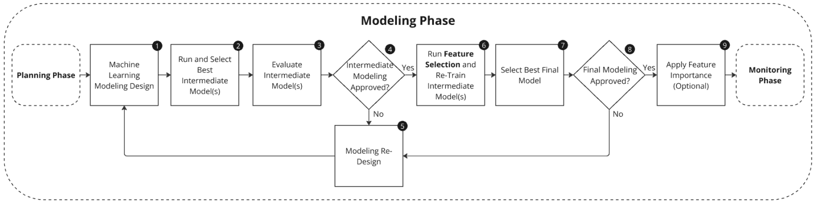

Figure 4: Modeling Phase Flowchart.

The **Monitoring phase** covers model implementation and corrective actions (it is **not** addressed in this research).

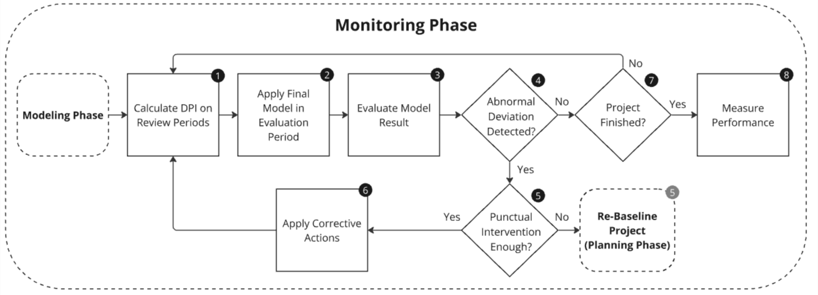

Figure 5: Monitoring Phase Flowchart.

Project Description and Monte Carlo Simulation

Project Description and Parameters

Table 1: Project Network and PDF Parameters

Activity, i	Predecessor	Min, a	ML, c	Max, b	PD_i	σ_i^2
1 - Engineering						
2		10	15	20	15	4.2
3	2	25	30	35	30	4.2
4	2	20	30	40	30	16.7
5	3, 4	45	55	65	55	16.7
6	3	60	70	80	70	16.7
7	3	80	90	100	90	16.7
8	2	50	70	90	70	66.7
9 - Procurement						
10	6	20	25	30	25	4.2
11	6	70	85	100	85	37.5
12	7	70	85	100	85	37.5
13	8	70	80	90	80	16.7
14	8	100	110	120	110	16.7
15	8	70	80	90	80	16.7
16	13, 15	25	30	35	30	4.2
17	16	12	15	18	15	1.5
18	2	170	190	210	190	66.7
19 - Construction						
20	3	45	55	65	55	16.7
21	20	50	60	70	60	16.7
22	4	45	55	65	55	16.7
23	5, 22	80	95	110	95	37.5
24	21	20	25	30	25	4.2
25	24	18	20	22	20	0.7
26	21	35	40	45	40	4.2
27	24	30	40	50	40	16.7
28	10	35	45	55	45	16.7
29	24, 28	75	85	95	85	16.7
30	11, 25, 26	75	80	85	80	4.2
31	12, 27	45	60	75	60	37.5
32	14, 23	12	15	18	15	1.5
33	32, 17, 18	50	60	70	60	16.7
34	29, 33	12	15	18	15	1.5
35	34, 30, 31	12	15	18	15	1.5
36	35					

Min = minimum; ML = most likely; Max = Maximum; PD_i = Planned Duration (duration average) of i; σ_i^2 = Variance of i

Project Description:

A real-world South American civil project.

A long-term composed of 32 activities over more than a year.

A high parallel level (serial-parallel index = 0.29),

A total planned duration of 1825 days,

A baseline planned duration of 300 days.

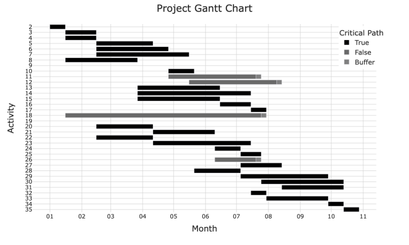

Figure 6: Gantt Chart Diagram.

DPI Analysis Over Time

At the beginning of the project, the DPI values **vary widely but converge over time**.

Their irregular distribution suggests that **fixed thresholds** (e.g., 0.8) are **unreliable** for detecting delays.

Although on-time and late projects show similar DPI ranges, their medians diverge progressively, indicating growing differences as the project advances.

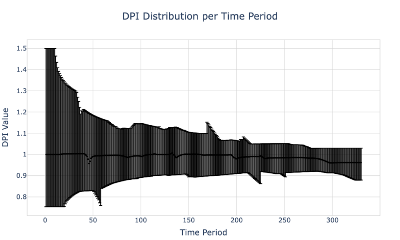

Figure 7: DPI Distribution per Time Period

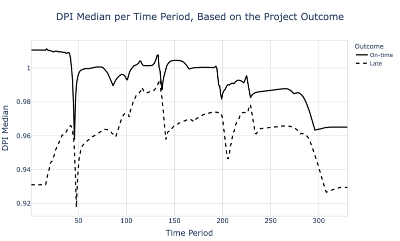

Figure 8: DPI Median per Time Period, Based on the Project Outcome

DPI Autocorrelation

As a cumulative metric, DPI is **strongly influenced** by past values.

Figure 9 shows high autocorrelation, especially at shorter lags:

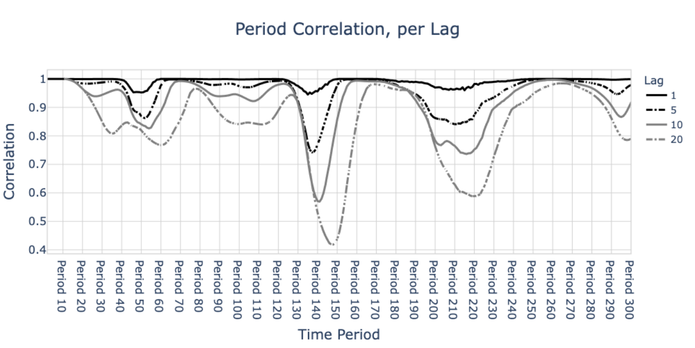

Figure 9: Correlation of the DPI Values.

Modeling Design Settings

Preprocessing Factors

Table 2 provides a summary of each **preprocessing technique** along with its corresponding levels:

Table 2: Preprocessing techniques and analysis factors.

Preprocessing Technique Factors	Labels of the Levels	# h_i
H_1: Missing values and outliers	None	1
H_2: Correlation removal	None* Remove 0.9 correlation	2
H_3: Dimensionality reduction	None* FA (keeping 0.95 of PCA variance)	2
H_4: Manual feature selection	$\mathbf{B}_1$* $\mathbf{B}_2$ $\mathbf{B}_3$ $\mathbf{B}_4$	4
H_5: Feature engineering	None* $\mathbf{D}_1$ $\mathbf{D}_2$	3
H_6: Data scaling	Standardization	1
H_7: Rebalancing strategy	None* Under-sampling Over-sampling	3
H_8: Train-test split	70% for training & 30% for testing	1

Table 3 provides a summary of the **ML modeling techniques** with its corresponding levels:

Table 3: Machine learning modeling and analysis factors.

ML modeling factors	Levels	# h_i
H_9: Performance metric	Area under the ROC curve (AUROC)	1
H_{10}: Hyperparameter tuning and search methods	Bayesian optimization	1
H_{11}: Cross-validation	5-fold cross validation	1
H_{12}: Machine learning models	Decision Tree (DT)* Random Forest (RF) XGBoost (XGB) LightGBM (LGBM)	4

Modeling Results

Outline

The **evaluation periods** were defined as $\mathbf{C} = \{100, 200, 300\}$.

Across all preprocessing and modeling levels, we tested 576 combinations per period, totaling 1,728 configurations.

With 5-fold cross-validation and 60 Bayesian optimization iterations, this resulted in approximately 518,400 trained ML models.

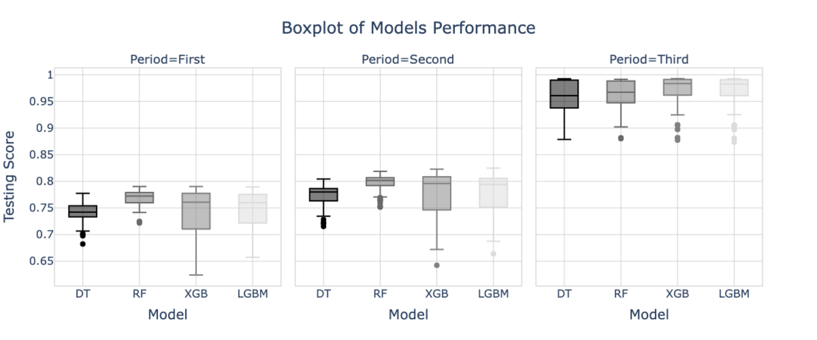

Figure 10: Boxplot of Models Performance

Outline

Intermediate and Final Models

Intermediate Models

Table 4: Intermediate Models and Preprocessing Techniques

Preprocessing and Modeling Factors	Evaluation Period		
	First ($t = (0, 100]$)	Second ($t = (100, 200]$)	Third ($t = (200, 300]$)
H_{12}: ML model	XGB	LGBM	RF
H_4: Manual feature selection	$\mathbf{B}_3^1 = \{10, 20 \ldots, 100\}$	$\mathbf{B}_3^2 = \{110, 120, \ldots, 200\}$	$\mathbf{B}_4^3 = \{220, 240, \ldots, 300\}$

Intermediate and Final Models Comparison

Table 5: Evaluation Period Results.

Evaluation Period	Model Version	Test Score (AUROC)	Review Periods
First ($t = (0, 100]$)	Intermediate	0.781	$\mathbf{M}_3^1 = \{20, 40, 50, 80, 100\}$
First ($t = (0, 100]$)	Final	0.779	$\mathbf{M}_3^1 = \{20, 50, 80, 100\}$
Second ($t = (100, 200]$)	Intermediate	0.810	$\mathbf{M}_3^2 = \{110, 140, 150, 200\}$
Second ($t = (100, 200]$)	Final	0.809	$\mathbf{M}_3^{2*} = \{150, 200\}$
Third ($t = (200, 300]$)	Intermediate	0.991	$\mathbf{M}_4^3 = \{240, 300\}$
Third ($t = (200, 300]$)	Final	0.991	$\mathbf{M}_4^{3*} = \{240, 300\}$

The total number of selected features was reduced from **25 to 8** (**68%** overall reduction), while maintaining high predictive performance across all evaluation periods.

- **First Period ($t = (0, 100]$):**
  - Intermediate model: 5 review periods, test score $=$ **0.781**
  - Final model: $\mathbf{M}_3^{1*} = \{20, 50, 80, 100\}$ (removed $t = 40$)
  - **60%** reduction in selected feature from original model
  - **0.002** (0.26%) performance loss from intermediate to final model
- **Second Period ($t = (100, 200]$):**
  - Intermediate model: 4 review periods, test score $=$ **0.810**
  - Final model: $\mathbf{M}_3^{2*} = \{150, 200\}$ (removed $t = 110, 140$)
  - **80%** reduction in selected features from original model
  - **0.001** (0.12%) performance loss from intermediate to final model
- **Third Period ($t = (200, 300]$):**
  - Intermediate model: 2 review periods, test score $=$ **0.991**
  - Final model: $\mathbf{M}_4^{3*} = \{240, 300\}$ (same as intermediate)
  - **60%** reduction in selected features from original model
  - **0** performance loss from intermediate to final model

Outline

ROC Curve

The ROC curve shows:

- Second period slightly outperforms the first (AUROC of 0.81 vs. 0.78)
- Third period has an AUROC of 0.99, indicating excellent performance.

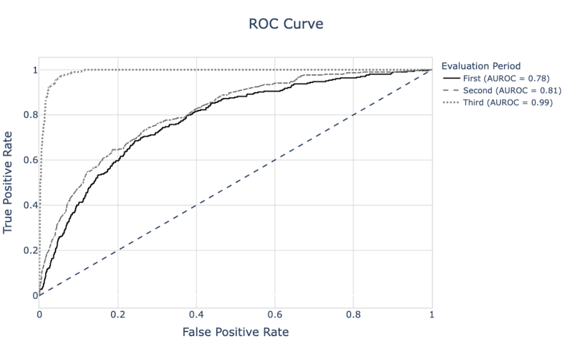

Figure 11: Receiver Operator Characteristic (ROC) Curve.

The first two evaluation periods show **similar patterns**, with most values below 50%. The second period is more dispersed, with on-time classifications concentrated near 0. In contrast, the third period shows **clear class separation**: on-time projects cluster near 0, while late ones are mostly above 0.7.

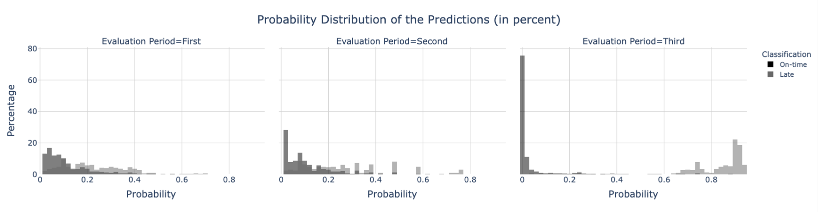

Figure 12: Probability Distribution Histogram of the Evaluation Period Predictions (in Percentage).

Final Remarks

Contributions, Limitations, and Next Steps

Research Contributions

- Reduced review periods **from 25 to 8** with less than **0.5% loss** in predictive performance.
- **Random Forest** with **manual feature selection** and **feature engineering** improved performance.
- **Oversampling** and **dimensionality reduction** decreased performance.

Limitations

- **Single-case scope:** Based on one project with long duration, high cost, and high parallelism → limits generalizability.
- **Model diversity:** Focused only on **tree-based models** (RF, XGBoost, LightGBM); other approaches not explored and could provide different insights.

Final Remarks and Next Steps

- Rather than offering definitive answers, the proposed tools and approaches serve as **decision-support systems** for managers.
- There is an inherent **trade off between false alarms and missed warnings**. A false positive may prompt unnecessary investigations, consuming time and resources. A false negative, on the other hand, may result in missed deadlines, financial penalties, or damage to organizational credibility.

References

Kursa, M. B. & Rudnicki, W. R. (2010), 'Feature selection with the Boruta package', *Journal of statistical software* **36**, 1–13.

Vanhoucke, M. (2023), 'The illusion of control', *Management for Professionals* .

Votto, R., Lee Ho, L. & Berssaneti, F. (2020), 'Applying and assessing performance of earned duration management control charts for epc project duration monitoring', *Journal of Construction Engineering and Management* **146**(3), 04020001.