Forecasting Notarial Stamp Revenue Using Machine Learning: A Comparative Study at the Court of Justice of Goiás

Renato Aparecido dos Santos Júnior[1], **Leonardo Afonso Amorim**[1], **Rafael Teixeira Sousa**[2]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG) – Goiânia, GO – Brazil

[2]Universidade Federal de Mato Grosso (UFMT) – Cuiabá, MT – Brazil

renatoasjunior@hotmail.com, leonardoafonso@gmail.com, rafaelts777@gmail.com

Abstract. *The predictability of notarial stamp revenue constitutes a critical challenge for budget planning in the Brazilian Judiciary. This study uses 74 months of data from the Sistema Extrajudicial Eletrônico (SEE), the electronic system used by the Court of Justice of Goiás to record and monitor notarial stamp transactions and revenue, covering Nov. 2019 to Dec. 2025, with IPCA as an exogenous variable, to compare statistical time-series and machine-learning models, including ARIMA, Prophet, XGBoost, LightGBM, hybrid models, and ensembles, under one-step-ahead and multi-step forecasting regimes. The multiplicative hybrid XGBoost achieved the best performance in the one-step setting (MAPE 3.14%, R^2 = 0.738), while the weighted ensemble stood out in the multi-step setting (MAPE 3.04%, R^2 = 0.770). As a practical contribution, an interactive Streamlit prototype was developed to support institutional decision-making in the TJGO.*

1. Introduction

Efficient budget management is one of the pillars of contemporary public administration, particularly in the context of the Brazilian Judiciary, whose operational capacity depends directly on the predictability of collected revenues. At the Court of Justice of Goiás (TJGO), revenue from notarial stamps, derived from fees and transfers to the Fund for the Development of the Judicial System (FUNDESP), represents a substantial share of the institutional budget. This revenue increased from levels close to R$ 8 million per month in 2019 to amounts exceeding R$ 20 million in 2025, according to data from the Sistema Extrajudicial Eletrônico (SEE), a tool of the Corregedoria-Geral da Justiça (CGJGO) used in this study.

Despite its financial relevance, the process of forecasting notarial revenue at the TJGO still relies on traditional forecasting practices, such as manual projections based on simple historical averages, without adequately incorporating seasonality, trends, and external variables. This limitation produces informational lag and reduces the institution's ability to respond to abrupt decreases or increases in revenue collection.

The problem analyzed presents characteristics that are favorable to the application of advanced predictive models. The availability of monthly historical series and relevant external variables, such as inflation indicators and exogenous attributes related to the economic context, makes this setting suitable for both classical decomposition methods [Box et al. 2015] and modern machine learning approaches for time series [Hyndman and Athanasopoulos 2018].

In this context, this study comparatively investigates predictive approaches applied to the TJGO notarial stamp revenue series. The statistical ARIMA model was evaluated as a baseline, along with additive decomposition models (Prophet), gradient boosting models based on decision trees (XGBoost and LightGBM), hybrid architectures, and a weighted ensemble. The integration of these models into the budget forecasting of notarial stamp revenues remains relatively underexplored in the scientific literature, which has predominantly focused on higher-volume federal and state taxes [Makridakis et al. 2022]; [Cerqueira et al. 2020]. From a practical standpoint, gains in predictive accuracy may reduce budget uncertainty, improve the planning of personnel and resources, and enable earlier detection of anomalies in revenue collection.

Accordingly, the general objective of this research is to investigate and compare the effectiveness of statistical time series models and machine learning models in forecasting recurrent organic notarial stamp revenue at the TJGO. Specifically, the study seeks to analyze seasonal patterns, trends, and outliers in the historical series; evaluate and compare predictive performance; investigate the effects of exogenous and calendar variables; and identify the most appropriate approach for institutional decision support.

1.1. Contributions of the Study

This study makes four main contributions. The first consists of a systematic comparison among ARIMA, Prophet, XGBoost, LightGBM, and hybrid architectures applied to notarial revenue forecasting in an institutional context that remains underexplored. The second lies in the comparative evaluation of one-step-ahead and multi-step forecasting regimes, preserving fair comparison across models and temporal validity. The third contribution concerns the investigation of the behavior of hybrid architectures, with particular emphasis on the multiplicative formulation, as an alternative for dealing with instability and drift in institutional time series. The fourth contribution consists of the development and validation of a predictive pipeline tailored to the TJGO context, demonstrating the practical applicability of the proposed approach in a real-world institutional environment.

2. Related Work

2.1. Time series forecasting

Time series are sequences of observations ordered in time, where the dependency between past and future values underpins the forecasting process [Box et al. 2015, Hyndman and Athanasopoulos 2018]. Modeling requires the decomposition of the series into trend, seasonality, and irregular components, alongside the analysis of stationarity and autocorrelation (ACF/PACF) to identify underlying patterns [Box et al. 2015, Hyndman and Athanasopoulos 2018]. In contexts such as public revenue collection, the dynamics are frequently affected by exogenous shocks and structural breaks, necessitating robust models [Makridakis et al. 2022]. Recent reviews also highlight the continued relevance of comparative studies involving classical statistical, machine learning, and hybrid forecasting approaches [Houssein et al. 2025, Sina et al. 2023].

For forecasting, the literature distinguishes between one-step-ahead, recursive multi-step, and direct strategies [Bontempi et al. 2013, Ben Taieb et al. 2012]. The ARIMA model is notable for its theoretical soundness and interpretability

[Box et al. 2015, Morettin and Toloi 2006], while Prophet offers flexibility in capturing seasonalities and external regressors [Taylor and Letham 2018]. Alternatively, gradient boosting algorithms such as XGBoost and LightGBM [Chen and Guestrin 2016, Ke et al. 2017] require explicit feature engineering to represent the temporal structure [Cerqueira et al. 2020]. The literature suggests that hybrid models and ensembles can mitigate individual limitations by combining the parsimony of linear models with the non-linear capacity of machine learning algorithms to enhance predictive robustness [Zhang 2003, Wolpert 1992, Zhou 2012, Sina et al. 2023].

2.2. Temporal evaluation

The evaluation of temporal models must preserve chronological order to avoid data leakage, employing protocols such as walk-forward validation and historical backtesting to ensure stability and robustness [Hyndman and Athanasopoulos 2018, Cerqueira et al. 2020, Bergmeir et al. 2018]. Predictive quality is assessed through complementary metrics, such as MAE, RMSE, MAPE, R^2, and MDA [Hyndman and Koehler 2006, Botchkarev 2019]. Additionally, residual diagnostics are essential to verify the absence of unexplained structure, using tests such as Ljung-Box, Durbin-Watson, and Shapiro-Wilk [Box et al. 2015, Shapiro and Wilk 1965]. Finally, the Diebold-Mariano test is employed to determine whether differences in accuracy between competing models are statistically significant [Diebold and Mariano 1995], with this methodological rigor being fundamental for addressing short series subject to institutional instabilities.

3. Methodology

3.1. Research design and methodological framework

This research is characterized as applied, quantitative, predictive, and comparative. The objective is to develop and evaluate models for forecasting the monthly revenue of notary stamps, utilizing statistical, machine learning, and hybrid approaches. The methodology adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) as a reference, adapted for time series [Wirth and Hipp 2000]. The stages followed were: business and data understanding, preparation, modeling, evaluation, and deployment. The following subsections detail each phase.

3.2. Empirical context, computational environment, and data sources

The study focuses on the monthly revenue of notary stamps from the Court of Justice of the State of Goiás (TJGO). The unit of analysis is the aggregated monthly time series, representing the total revenue per month. All stages of processing, feature engineering, modeling, and evaluation were performed in Python, in a Jupyter Notebook environment, using libraries such as pandas, numpy, matplotlib, seaborn, statsmodels, Optuna, and python-bcb. This approach ensured traceability, organization of experiments, and reproducibility of the methodological workflow, integrating exploration, construction of the analytical base, temporal partitioning, modeling, and comparison of results in a single workspace. The data were extracted from the TJGO SEE system, complemented by the IPCA (IBGE/Central Bank). The raw period covers November 2019 to December 2025 (74 months).

3.3. Data preparation, exploratory analysis, and feature engineering

The target variable is the recurring organic revenue, composed of Fundesp and judicial fees, excluding interim surplus and punitive or late-payment revenues. The first 13 months (burn-in) were discarded for series stabilization, resulting in 61 observations. Figure 1 presents the complete time series, showing a growth trend and annual seasonality. Seasonal decomposition in Figure 2 confirmed trend, seasonality, and residual components.



Figure 1. Exploratory analysis of the TJGO revenue series.

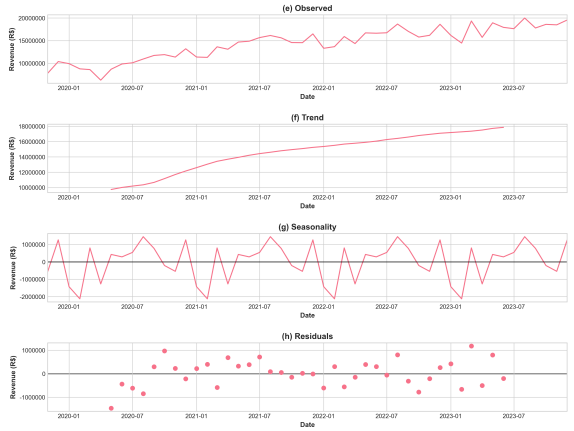


Figure 2. Seasonal decomposition of TJGO revenue in the training set.

A comprehensive feature engineering process was conducted to represent temporal memory, seasonality, trend, and macroeconomic effects. The predictor set included lagged variables, rolling-window statistics, calendar-based features (month and quarter), cyclic seasonal encodings using sine and cosine transformations, and the IPCA inflation index as an exogenous variable. The selection of lag structures was guided by the autocorrelation (ACF) and partial autocorrelation (PACF) analyses. Figure 3 illustrates the temporal dependencies observed in the series and the evidence used to define the lag configuration adopted in the models. To avoid data leakage and preserve the validity of the forecasting exercise, all predictors were derived exclusively from information available at prediction time. Consequently, lag 1 was excluded, rolling statistics were shifted, and threshold-based variables were computed using fixed parameters estimated from the training period.

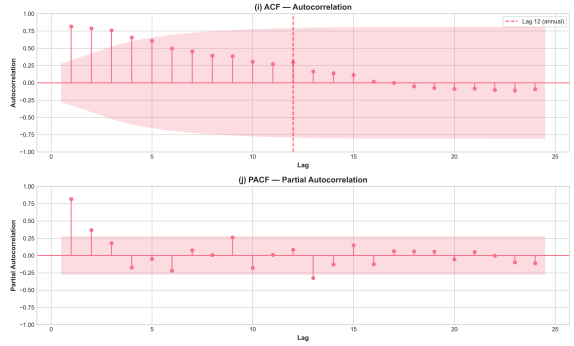


Figure 3. ACF and PACF functions of the revenue series in the training set.

3.4. Predictive modeling, temporal validation, and model tuning

The base was partitioned into training (Dec/2020 to Dec/2023), validation (2024), and test (2025). Validation adopted walk-forward with an expanding window, re-estimating the models at each step. Three regimes were evaluated: one-step-ahead (1 month ahead), recursive multi-step (12 months ahead, using predicted values as inputs for subsequent steps), and direct multi-step (12 months ahead, with one model per horizon). The implemented models were ARIMA baseline, Prophet baseline (seasonality only), Prophet with regressors, LightGBM baseline, XGBoost baseline, additive hybrid, regularized additive hybrid (LASSO), multiplicative hybrid, and weighted ensembles. The additive hybrid model combines the Prophet forecast with residual correction via XGBoost: $\hat{y}_t = \hat{p}_t + \hat{r}_t$, whereas the multiplicative hybrid is given by $\hat{y}_t = \hat{p}_t \times (1 + \hat{r}_t)$.

The Ensemble calculates a weighted average of individual forecasts, with weights defined by the inverse of the validation-set MAPE, so lower-error models contribute more to the final forecast. The composition is deterministic and reproducible. Its components were selected based on methodological complementarity, combining distinct predictive structures in the same arrangement. To handle structural changes, binary variables indicating the months of the COVID-19 pandemic (March to December 2020) and atypical shocks identified in the series were incorporated, while IPCA was included as an exogenous variable to capture inflationary effects on revenue. Table 1 summarizes the predictor variables used, including lags, moving averages, calendar, cyclic, IPCA, and event markers. Model hyperparameters were optimized with Optuna, using the TPE (Tree-structured Parzen Estimator) algorithm and walk-forward validation as the objective function.

3.5. Evaluation metrics, comparative testing, and residual diagnostics

The main metrics were MAPE, sMAPE, MAE, RMSE, R^2, and MDA. For $e_t = y_t - \hat{y}_t$, $\text{MAE} = \frac{1}{n}\sum|e_t|$, $\text{RMSE} = \left(\frac{1}{n}\sum e_t^2\right)^{1/2}$, $\text{MAPE} = \frac{100}{n}\sum\left|\frac{e_t}{y_t}\right|$, $\text{sMAPE} = \frac{200}{n}\sum\frac{|e_t|}{|y_t|+|\hat{y}_t|}$, $R^2 = 1 - \frac{\sum e_t^2}{\sum(y_t-\bar{y})^2}$, and $\text{MDA} = \frac{100}{n-1}\sum_{t=2}^{n} I[sign(y_t - y_{t-1}) = sign(\hat{y}_t - \hat{y}_{t-1})]$. Model comparisons used the Diebold-Mariano test, and residuals were assessed with Ljung-Box, Durbin-Watson, and Shapiro-Wilk tests.

4. Forecasting Architecture

The proposed forecasting architecture is organized into five layers: data ingestion, analytical preparation, predictive modeling, model evaluation, and decision support. Figure 4

Table 1. Groups of predictor variables and their functions in modeling

Category	Examples of variables	Function
Calendar and seasonality	quarter, business_days, month_sin, month_cos	Capture seasonal patterns and calendar effects
Lags and variations	lag_2, lag_3, monthly_variation_pct_lag1, avg_ticket_lag1	Model temporal dependence and lagged effects
Shocks and thresholds	pandemic, statistical_ipca_shock, statistical_ipca_threshold	Identify structural breaks and non-linear effects
Macroeconomic and institutional	ipca_variation, ipca_accum_12m, n_notaries	Control macroeconomic and institutional context
Rolling window and trend	moving_avg_3m, moving_avg_12m, moving_std_3m, trend	Extract short and long-term trends and volatility

summarizes the workflow, illustrating the flow from data acquisition and feature engineering to forecasting, validation, and operational deployment.

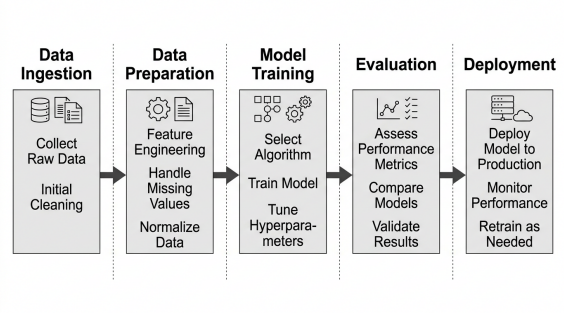


Figure 4. Architecture of the proposed predictive solution for the TJGO.

5. Results and Discussion

5.1. Model Performance Comparison

The empirical evaluation of the proposed solution was conducted based on strict chronological partitioning, preserving the separation between training, validation, and testing sets to replicate conditions consistent with prospective use. Twenty models were compared, encompassing formulations based on ARIMA, LightGBM, Prophet, XGBoost, hybrid architectures, and ensemble strategies under different forecasting regimes. Table 2 presents the consolidated comparison of the models, including metrics for percentage error, monetary scale error, goodness-of-fit, and directional accuracy.

In the analyzed dataset, multiplicative hybrid approaches and ensemble strategies yielded the best predictive performance. The best overall performance on the test set was achieved by the Ensemble in the Multi-Step regime, which presented a MAPE of 3.044%, sMAPE of 3.046%, MAE of R$ 659,342.00, RMSE of R$ 758,225.00, $R^2 = 0.770$, and MDA of 90.909%. The Multiplicative XGBoost models also stood out in both regimes, suggesting that the combination of explicit temporal structure with residual correction better captured proportional variations and local deviations in the series. In contrast, the XGB Baseline formulations, without a hybrid structure, presented substantially inferior

Table 2. Comparative performance of predictive models

Model	Regime	MAPE Val (%)	MAPE Test (%)	sMAPE Test (%)	MAE Test (R$)	RMSE Test (R$)	R^2 Test	MDA Test (%)
Ensemble	Multi-Step	2.389	3.044	3.046	659,342.00	758,225.00	0.770	90.909
XGB Multiplicative	One-Step	2.596	3.140	3.128	678,351.00	808,358.00	0.739	90.909
XGB Multiplicative	Multi-Step	2.617	3.164	3.150	682,816.00	809,264.00	0.738	90.909
Ensemble	Mixed	2.322	3.297	3.294	711,597.00	828,874.00	0.725	90.909
Ensemble	One-Step	2.343	3.608	3.598	775,797.00	901,785.00	0.675	90.909
Prophet + Regressors	Multi-Step	3.896	3.631	3.723	803,564.00	999,661.00	0.601	90.909
XGB Additive L1-L2	Multi-Step	3.913	4.239	4.148	894,007.00	1,021,955.00	0.583	72.727
Prophet + Regressors	One-Step	4.215	4.258	4.279	912,867.00	1,005,005.00	0.596	81.818
XGB Additive	Multi-Step	3.909	4.470	4.370	941,511.00	1,065,361.00	0.546	81.818
XGB Additive L1-L2	One-Step	3.232	4.490	4.453	960,406.00	1,107,549.00	0.510	81.818
XGB Additive	One-Step	3.412	4.657	4.607	992,936.00	1,136,325.00	0.484	81.818
XGB Additive Multi-Direct	Multi-Step	3.408	4.901	4.817	1,036,274.00	1,148,520.00	0.473	90.909
Prophet Baseline	One-Step	4.510	5.413	5.441	1,166,228.00	1,349,535.00	0.272	54.545
Prophet Baseline	Multi-Step	3.735	5.704	5.857	1,251,654.00	1,524,209.00	0.071	27.273
ARIMA Baseline	One-Step	5.223	6.053	6.097	1,328,072.00	1,522,621.00	0.073	27.273
LightGBM Baseline	One-Step	8.273	8.237	8.725	1,833,924.00	2,163,425.00	-0.871	72.727
LightGBM Baseline	Multi-Step	17.610	9.329	10.007	2,101,365.00	2,584,751.00	-1.670	27.273
ARIMA Baseline	Multi-Step	5.650	11.061	11.926	2,494,188.00	2,918,769.00	-2.405	18.182
XGB Baseline	Multi-Step	5.446	11.588	12.482	2,598,150.00	2,949,534.00	-2.477	90.909
XGB Baseline	One-Step	7.088	15.567	17.047	3,453,085.00	3,707,941.00	-4.495	72.727

results, with high MAPE and negative R^2, indicating lower temporal extrapolation capacity. Several models achieved MDA = 90.909%, corresponding to 10 correct directional predictions out of 11 month-to-month comparisons in the 12-month test period; because MDA is discrete and based only on direction, distinct models may share the same value despite different forecast magnitudes.

Although metrics such as MAE, RMSE, and MAPE allow for comparing predictive performance between models, their interpretation becomes clearer when translated to the scale of the problem. In the best model evaluated, the mean absolute error of approximately R$ 659 thousand corresponds to about 3% of a monthly revenue exceeding R$ 20 million, indicating that, despite the oscillations inherent to the series, the observed error level remains relatively low in operational terms. From a managerial perspective, this suggests that the proposed solution can provide useful support for budget planning, resource provisioning, and monitoring revenue deviations with a degree of precision compatible with the institutional context analyzed.

Compared with the ARIMA baseline, the Ensemble Multi-Step reduced forecasting error by 72.5%, while the Multiplicative XGB achieved reductions of 71.4% and 48.1% in the Multi-Step and One-Step regimes, respectively. These results reinforce the contribution of hybrid and ensemble architectures to institutional revenue forecasting.

The weights assigned to the Ensemble components, derived from validation-set performance, help explain this result. Prophet + Regressors contributed 28.70% of the final forecast, preserving trend, seasonality, and exogenous information. The Multiplicative Hybrid XGB received the largest weight (42.72%), indicating greater usefulness in modeling non-linear and proportional fluctuations not fully captured by the additive structure. The L1/L2 Additive Hybrid XGB accounted for 28.58% and acted as a stabilizing

component, adding regularization and reducing overfitting risk. Taken together, these results suggest that the Ensemble performed better because it combined complementary mechanisms rather than relying on a single predictive structure.

Figure 5 complements the analysis by presenting, in a single panel, a visual comparison between the observed series and the forecasts of the seven top-ranked models. The graph demonstrates that the leading approaches more closely track the actual trajectory of the revenue.

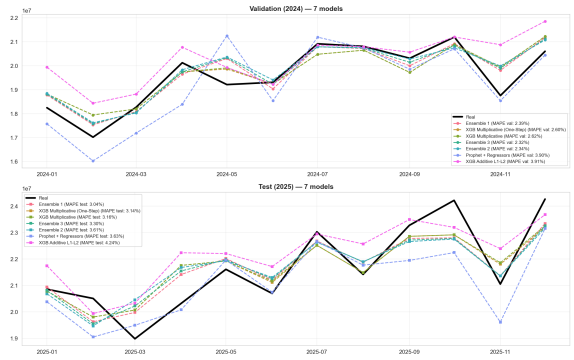


Figure 5. Observed series and forecasts of the seven best-ranked models.

5.2. Residual diagnostics and statistical comparison

Table 3 presents the statistical diagnostics of the residuals in the test set. The best-performing models exhibited residual behavior more consistent with the absence of remaining temporal structure, with Durbin-Watson statistics close to the desirable range and no strong evidence of residual autocorrelation. Conversely, the baseline XGBoost formulations showed less favorable signs, suggesting that a relevant portion of the temporal dynamics remained uncaptured.

To complement the comparison between competing forecasts, the Diebold-Mariano test was applied, taking the Ensemble in the Multi-Step regime as a reference. The results (Table 4) show a statistically significant difference relative to some competitors and a statistical tie with the most competitive formulations. In this context, the Diebold-Mariano test offers additional statistical support for the comparative ranking, although the small test sample limits its power to detect subtle differences in predictive accuracy.

Residual diagnostics and comparative testing reinforce the consistency of the selected solution, demonstrating that its performance stems from stable behavior across different evaluation criteria.

5.3. Backtesting Analysis

The analysis was complemented by expansive backtesting to examine the performance evolution of the six main models across different time windows. Figure 6 summarizes this temporal comparison: hybrid and ensemble models showed a progressive reduction in errors as the historical data expanded, whereas simpler models revealed greater sensitivity in the initial segments. This result reduces the risk of interpretation based solely on a single test cut and suggests that the proposed solution is well-suited to the dynamic nature of the series.

Table 3. Model residual diagnostic tests

Model	Regime	DW Test	Shapiro Test (p)	LB6 Test (p)
Ensemble	Multi-Step	1.822	0.645	0.369
XGB Multiplicative	One-Step	1.875	0.988	0.612
XGB Multiplicative	Multi-Step	1.916	0.988	0.586
Ensemble	Mixed	1.774	0.703	0.388
Ensemble	One-Step	1.736	0.774	0.422
Prophet + Regressors	Multi-Step	0.845	0.338	0.143
XGB Additive L1-L2	Multi-Step	1.790	0.634	0.203
Prophet + Regressors	One-Step	1.548	0.350	0.453
XGB Additive	Multi-Step	1.658	0.412	0.255
XGB Additive L1-L2	One-Step	1.813	0.463	0.565
XGB Additive	One-Step	1.702	0.482	0.556
XGB Additive Multi-Direct	Multi-Step	1.730	0.200	0.559
Prophet Baseline	One-Step	1.895	0.633	0.548
Prophet Baseline	Multi-Step	1.699	0.987	0.556
ARIMA Baseline	One-Step	2.649	0.323	0.065
ARIMA Baseline	Multi-Step	0.385	0.371	0.133
XGB Baseline	Multi-Step	0.259	0.361	0.059
XGB Baseline	One-Step	0.151	0.982	0.088
LightGBM Baseline	One-Step	0.619	0.492	0.705
LightGBM Baseline	Multi-Step	0.560	0.128	0.496

Table 4. Diebold-Mariano test against Ensemble

Model	DM Stat	p-value	Result
Prophet Baseline	2.1483	0.0317	Significant difference
ARIMA Baseline	3.0479	0.0023	Significant difference
LightGBM Baseline	2.4659	0.0137	Significant difference
XGB Baseline	3.3391	0.0008	Significant difference
Prophet + Regressors	1.4405	0.1497	Statistical tie
XGB Additive	1.9719	0.0486	Significant difference
XGB Additive L1-L2	1.6318	0.1027	Statistical tie
XGB Multiplicative	0.8073	0.4195	Statistical tie
XGB Additive Multi-Direct	2.5121	0.0120	Significant difference

5.4. Importance of predictor variables

The relative importance of variables in the XGBoost-based models was examined. Lagged variables, moving averages, trend-related attributes, and calendar features emerged as the most influential predictors. Attributes associated with temporal memory, trend, operational calendar, and recent dynamics stand out. Calendar-related features reinforced the presence of seasonal patterns in the series, while exogenous variables such as IPCA played a complementary role, contributing to performance improvements when combined with temporal attributes.

5.5. Interpretation of temporal patterns and predictor variables

The exploratory and statistical analyses indicate that TJGO revenue is primarily driven by trend and seasonal components. The ACF and PACF analyses revealed short-term temporal dependence, while correlation analysis showed stronger associations with lagged variables and moving averages than with exogenous indicators such as IPCA. These find-

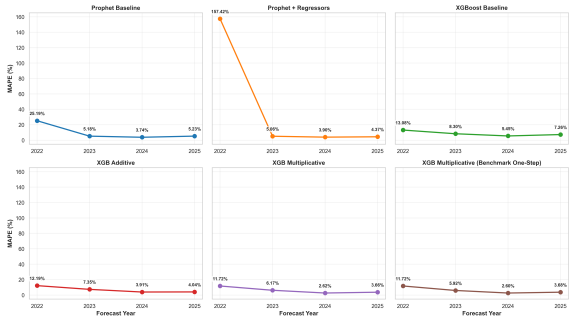


Figure 6. Learning Curves by Predictive Architecture.

ings are consistent with the superior performance of hybrid and ensemble architectures, which effectively combine temporal memory, trend information, and residual dynamics.

5.6. Streamlit-Based Decision Support Prototype

A Streamlit-based prototype was implemented to operationalize the proposed solution and support institutional decision-making. Figure 7 illustrates the resulting interface.

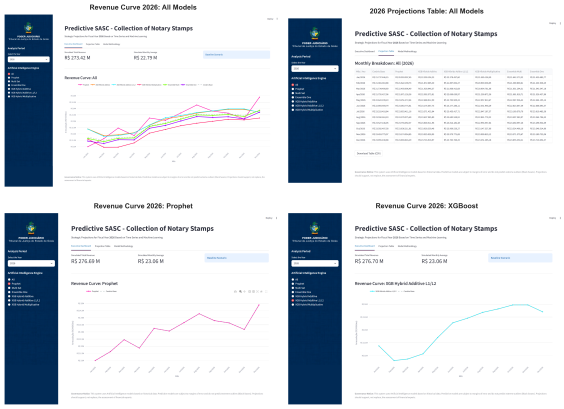


Figure 7. Streamlit-Based Decision Support Prototype.

6. Conclusion

This work developed and evaluated a predictive solution for the monthly revenue of the Court of Justice of the State of Goias, combining rigorous temporal processing, feature engineering, and comparison between different modeling approaches. The results showed that, in the analyzed context, approaches based on forecast combination and hybrid structures were more suitable, with the Ensemble Multi-Step model standing out as the best-performing specification. The combination of trend, temporal memory, contextual variables, and forecast aggregation produced superior performance compared to simpler specifications, with hybrid architectures and ensembles exhibiting more robust behavior. Temporal modeling with strict separation between training, validation, and testing was fundamental to avoid overestimation and to approximate the evaluation to real conditions of prospective use. The main contribution of the study is to offer a replicable methodological arrangement for revenue monitoring, comparison between competing approaches, and periodic updating of estimates, expanding the analytical use of institutional

data in public management. Furthermore, the reduction in forecasting error provides practical support for budget planning, resource allocation, and the early identification of revenue deviations, thereby strengthening institutional decision-making.

Among the limitations of this study, the historical series contains only 61 effective observations after the burn-in period, with 11 observations in the test set, which limits the robustness of model comparisons and the generalizability of the findings. Furthermore, the study was conducted using data from a single Brazilian court and a single revenue dataset, so additional validation is required before extending the conclusions to other institutional contexts. Although the adoption of a parsimonious set of exogenous variables favored methodological stability and reduced the risk of overfitting, other potentially relevant dimensions of revenue behavior were not explored in depth.

Future research should investigate the incorporation of additional exogenous variables, including indicators related to economic activity, credit markets, real estate transactions, and civil registry operations. It would also be valuable to expand the interpretability analysis of tree-based models and evaluate more recent forecasting architectures, such as Temporal Fusion Transformers (TFT) and N-HiTS. Additional research directions include the development of hierarchical forecasting models at the notary-office level, the integration of anomaly detection techniques, and the incorporation of economic nowcasting indicators. Finally, the evolution of the proposed solution into a continuous operational environment, with automated model retraining and systematic performance monitoring, represents a natural step toward consolidating its use as an institutional decision-support tool.

Overall, the results indicate that the proposed workflow can support data-driven budget planning in the institutional context analyzed.

References

Ben Taieb, S., Atiya, A. F., Bontempi, G., and Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications*, 39(8):7067–7083.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Bontempi, G., Ben Taieb, S., and Le Borgne, Y.-A. (2013). Machine learning strategies for time series forecasting. In *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures*, pages 62–77. Springer, Berlin; Heidelberg.

Botchkarev, A. (2019). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. arXiv:1809.03006. Disponïvel em: `https://arxiv.org/abs/1809.03006`. Acesso em: 20 abr. 2026.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. Wiley, Hoboken, 5 edition.

Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 109:1997–2028.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco. ACM.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Houssein, E. H., Mohamed, M., Younis, E. M. G., et al. (2025). Artificial intelligence and classical statistical models for time series forecasting: a comprehensive review. *Journal of Big Data*, 12:271.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts, Melbourne, 2 edition. Disponïvel em: `https://otexts.com/fpp2/`. Acesso em: 20 abr. 2026.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 3146–3154.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022). Statistical and machine learning forecasting methods: concerns and ways forward. *PLOS ONE*, 17(3):e0264810.

Morettin, P. A. and Toloi, C. M. C. (2006). *Análise de Séries Temporais*. ABE – Manual do Usuário, 2 edition.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Sina, L. B., Secco, C. A., Blazevic, M., and Nazemi, K. (2023). Hybrid forecasting methods—a systematic review. *Electronics*, 12(9):2019.

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC Press, Boca Raton.