Machine Learning Forecasting of U.S. Motor Vehicle Production

Alfredo Sosa

2026-03-22

1 Introduction

This memo documents the implementation of three machine learning models used to forecast U.S. motor vehicle production. The objective is to complement the econometric analysis in Sosa (2026) with predictive methods commonly used in applied macroeconomics and industry analytics.

The models approximate a counterfactual production path using observable macroeconomic and industry-specific inputs. The resulting predictions are used to construct an interactive Tableau dashboard that provides an intuitive visualization of production dynamics.

The analysis considers a standard train/test Random Forest, a walk-forward Random Forest, and a Gradient Boosting model. These approaches progressively increase the realism of the forecasting exercise, moving from static estimation to real-time prediction.

2 Data

The dataset consists of monthly time-series observations drawn primarily from the Federal Reserve Economic Data (FRED) database. These series capture key dimensions of industrial activity, input costs, labor market conditions, and monetary policy relevant to the U.S. motor vehicle sector. The construction of the dataset follows Sosa (2026), where the raw series are harmonized and combined into a unified panel suitable for empirical analysis.

The dependent variable is defined as:

$$Y_t = \text{Motor Vehicle Production}_t$$

This series corresponds to the FRED index IPG3361S, which measures industrial production in the motor vehicles and parts sector and serves as the target variable in all machine learning models.

The explanatory variables include total industrial production (INDPRO), the producer price index for steel (WPU101), manufacturing employment (MANEMP), and the federal funds rate (FEDFUNDS). These variables jointly capture aggregate demand conditions, input cost pressures, labor market dynamics, and monetary policy stance, all of which are central determinants of production in the automotive sector.

The feature vector used in the models is given by:

$$X_t = \left(\text{PPI}_t,\ \text{Employment}_t,\ r_t,\ Y_{t-1}, Y_{t-2}, Y_{t-3}, IP_{t-1}, IP_{t-2}\right)$$

where the variables correspond directly to the FRED series described above. The inclusion of lagged production terms captures persistence and adjustment dynamics, while lagged industrial production reflects broader macroeconomic momentum.

All series are converted to monthly frequency, aligned temporally, and merged into a single dataset. Lagged variables are constructed explicitly, and the resulting panel is sorted chronologically to support both standard machine learning estimation and real-time forecasting exercises.

3 Methodology

The objective of each model is to estimate a nonlinear mapping from observable covariates to production:

$$\hat{Y}_t = f(X_t)$$

where the function (f( )) is approximated using tree-based ensemble methods. Prediction error is defined as:

$$\varepsilon_t = Y_t - \hat{Y}_t$$

and model performance is evaluated using the mean absolute error:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T} |Y_t - \hat{Y}_t|$$

The first specification relies on a conventional train/test split in which the model is estimated on pre-2014 data and evaluated on subsequent observations. This approach provides a baseline measure of predictive performance but implicitly assumes stability in the relationship between inputs and output over time.

The second specification adopts a walk-forward forecasting framework in which the model is recursively re-estimated using only past information. Formally, predictions are generated according to:

$$\hat{Y}_t = f_t(X_t), \quad \text{where } f_t \text{ is estimated using data up to } t-1$$

This procedure closely mirrors real-time forecasting and eliminates look-ahead bias. At each period, the model is trained on all available historical data and then used to predict the next observation, generating a sequence of out-of-sample forecasts.

The third specification employs Gradient Boosting, which constructs predictions through a sequence of weak learners that iteratively reduce residual error:

$$\hat{Y}_t = \sum_{m=1}^{M} \gamma_m h_m(X_t)$$

This approach emphasizes predictive accuracy by capturing nonlinearities and interactions in the data more effectively than standard ensemble methods.

4 Results

Across all three models, predicted production closely tracks the observed series, indicating that the selected covariates capture key drivers of U.S. motor vehicle output. The results reveal strong persistence in production dynamics, largely driven by lagged values of the dependent variable, alongside meaningful contributions from macroeconomic conditions such as employment and interest rates.

The Gradient Boosting model produces the smoothest fit and the highest predictive accuracy, while the walk-forward specification provides the most credible representation of real-time forecasting performance. Because it respects the information set available at each point in time, the walk-forward approach yields a particularly useful counterfactual series for policy interpretation.

Overall, prediction errors remain small relative to the scale of production, suggesting that the models provide a reliable approximation of underlying production dynamics.

5 Contributions

This exercise extends the empirical framework of Sosa (2026) by incorporating machine learning methods alongside traditional econometric approaches. In doing so, it introduces a complementary predictive perspective that enhances the analysis of industrial production dynamics.

The results also provide a data-driven benchmark for evaluating deviations in production that may be associated with structural shocks, including trade policy interventions. In addition, the project demonstrates the integration of data engineering, machine learning, and visualization tools within a unified workflow, highlighting the practical applicability of the methods.

The use of walk-forward forecasting aligns with best practices in applied macroeconomics and strengthens the credibility of the empirical results.

From a policy perspective, deviations between predicted and observed production provide an intuitive way to identify periods in which standard macroeconomic fundamentals fail to fully explain realized outcomes. In particular, systematic gaps between the machine learning predictions and actual production may reflect the influence of external shocks, including trade policy interventions such as the Section 301 tariffs analyzed in Sosa (2026). Because the models are trained on core macroeconomic and industry variables, they generate a counterfactual benchmark that abstracts from policy-specific effects. As a result, periods in which observed production persistently underperforms relative to predicted values can be interpreted as consistent with negative supply chain disruptions or cost shocks associated with tariff exposure. This predictive framework therefore complements the causal estimates by providing a transparent and data-driven way to visualize the aggregate implications of trade policy in real time.

6 Interactive Dashboard

An interactive Tableau dashboard accompanies this analysis and provides a visual representation of the machine learning forecasts. The dashboard integrates the outputs of the three models and allows users to compare predicted and observed production over time, examine prediction errors, and explore how model performance evolves across periods.

The dashboard is available at:

https://public.tableau.com/app/profile/alfredo.sosa/viz/dashboard_ml_simple/Dashboard1

This interactive component complements the statistical analysis by translating model outputs into an accessible visual format. It also illustrates the end-to-end integration of FRED-based data construction, Python-based machine learning models, and modern visualization tools, reinforcing the practical relevance of the approach.

7 Conclusion

The three machine learning models provide a flexible and accurate framework for forecasting U.S. motor vehicle production. While the train/test specification offers a useful benchmark, the walk-forward approach delivers a more realistic representation of real-time forecasting, and Gradient Boosting achieves the highest predictive performance.

These models complement the causal analysis in Sosa (2026) by offering a predictive perspective on production dynamics, thereby enhancing both analytical depth and practical relevance.