

# Proyeccion ARIMA/ETS

Gabriel Gonzalo Ojeda Cárcamo

7/1/2026

## Executive Summary

This project implements a monthly demand forecasting model for industrial products operating in a highly volatile environment with limited historical data. Using 40 monthly observations per product and no exogenous variables, several classical time series models (ARIMA, ETS, TBATS) were evaluated as benchmarks.

To better capture abrupt demand fluctuations, a LightGBM model was developed with recent shock features and validated using walk-forward (expanding window) cross-validation. This approach reduced prediction error by approximately 50% compared to an ARIMA baseline, lowering MAPE from 30% to 15% aprox.

The final solution also provides quantile forecasts (P10–P50–P90), enabling uncertainty-aware planning and inventory decision-making. The project highlights how domain-informed feature engineering and robust temporal validation can significantly improve forecasting performance in short, noisy time series.

The forecasts support monthly production planning and inventory decisions for three industrial products. Over-forecasting increases holding costs, while under-forecasting leads to stockouts and lost sales. Therefore, probabilistic forecasts (P10–P50–P90) are required to quantify uncertainty.

```
library(readxl)
library(tidyverse)
library(tsibble)
library(fable)
library(fabletools)
library(lubridate)
library(dplyr)
library(Metrics)
library(feasts)
library(lightgbm)
library(zoo)
library(forecast)
library(ggplot2)
```

## Data and scope

- Frequency: Monthly
- Period: 2021–2024
- Observations: 40 per product
- Products analyzed: Three industrial products .

The analysis is based on 40 monthly observations per product, with no exogenous variables available. This constraint requires the use of parsimonious and robust modeling approaches capable of handling limited data and high volatility.

```
data_prd <- read_excel(
  "C:/Users/gonza/OneDrive/Documentos/Time series/DATA PRD.xlsx"
)

#Wrangling

data_prd <- data_prd %>%
  arrange(date) %>%                                # CLAVE
  mutate(Date = yearmonth(date)) %>%
  as_tsibble(index = Date)

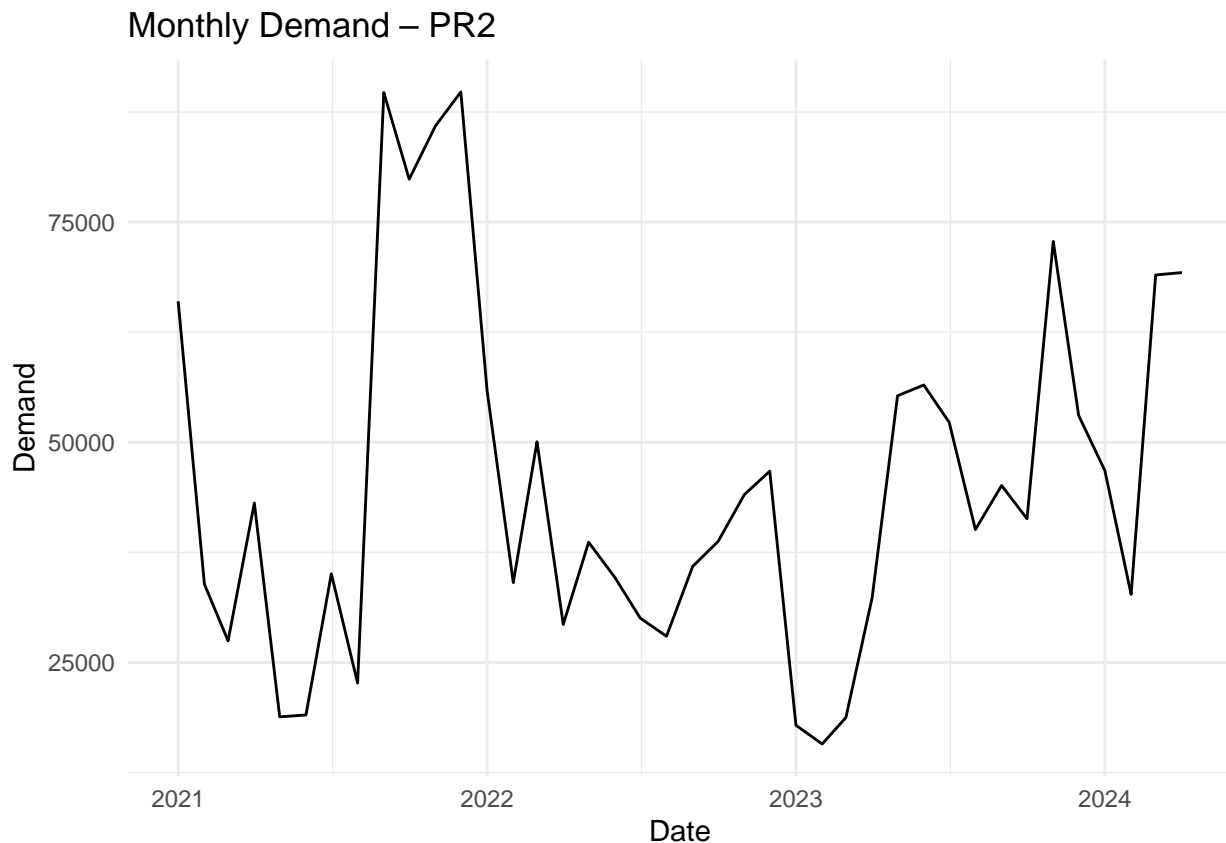
data_num <- data_prd %>%
  as_tibble() %>%                                    # rompe clases tsibble
  mutate(
    PR1 = as.numeric(PR1),
    PR2 = as.numeric(PR2),
    PR3 = as.numeric(PR3)
  ) %>%
  select(PR1, PR2, PR3)
```

## Exploratory Data Analysis

### Time series visualization

```
library(ggplot2)

ggplot(data_prd, aes(x = date, y = PR2)) +
  geom_line(color = "black") +
  theme_minimal() +
  labs(
    title = "Monthly Demand - PR2",
    x = "Date",
    y = "Demand"
  )
```



The time series shows a highly volatile demand pattern with abrupt spikes and sharp drops throughout the period 2021–2024. Demand levels fluctuate widely, ranging from periods of sustained low values to sudden peaks exceeding historical averages.

Several large, short-lived spikes are visible (notably around late 2021 and mid-2023), followed by rapid reversals. This behavior suggests that extreme values are driven by temporary shocks rather than persistent trend changes, making them difficult to capture using traditional smoothing-based models.

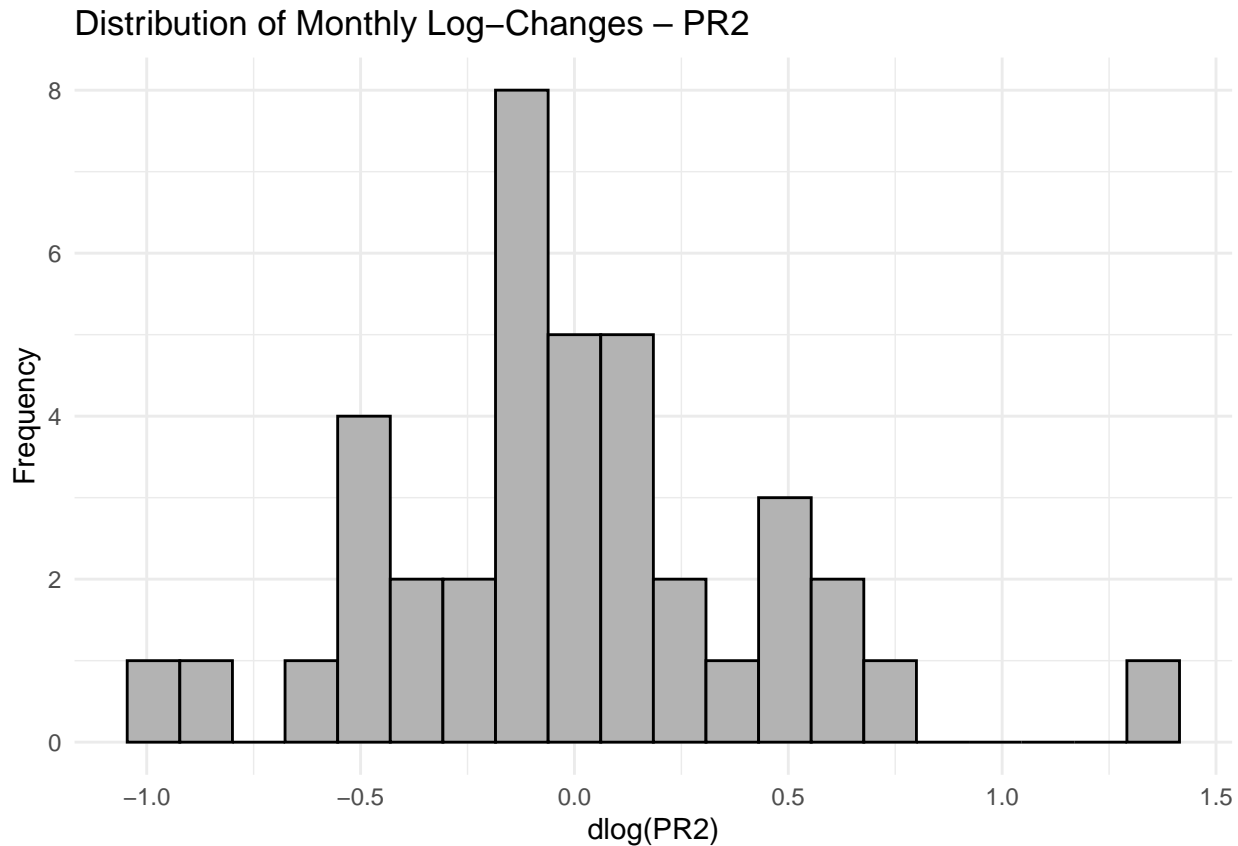
There is no clear, stable seasonal pattern across years. While some higher demand periods appear around year-end, the magnitude and timing of peaks vary significantly, indicating irregular seasonality rather than a fixed annual cycle.

## Monthly Change Distribution

```
data_prd %>%
  arrange(date) %>%
  mutate(dlog_PR2 = difference(log(PR2))) %>%
  ggplot(aes(x = dlog_PR2)) +
  geom_histogram(bins = 20, fill = "gray70", color = "black") +
  theme_minimal() +
  labs(
    title = "Distribution of Monthly Log-Changes - PR2",
    x = "dlog(PR2)",
    y = "Frequency"
  )
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
```

```
## ('stat_bin()').
```



The distribution of monthly log-changes shows asymmetry and heavy tails, indicating that demand variations are not normally distributed. This reinforces the need for models that can handle extreme movements and irregular behavior.

## Correlation between products

```
# Correlacion en niveles  
  
cor_levels <- cor(data_num, use = "complete.obs")  
round(cor_levels, 3)
```

```
##      PR1  PR2  PR3  
## PR1 1.000 0.672 0.731  
## PR2 0.672 1.000 0.953  
## PR3 0.731 0.953 1.000
```

While the products exhibit moderate to high correlation in levels, they are treated as independent forecasting problems, as no direct causal relationship exists. The observed co-movement is interpreted as a shared market effect rather than structural dependence, justifying univariate modeling.

## Change correlation

```
data_diff <- data_prd %>%
  as_tibble() %>%
  mutate(
    dPR1 = difference(log(as.numeric(PR1))),
    dPR2 = difference(log(as.numeric(PR2))),
    dPR3 = difference(log(as.numeric(PR3)))
  ) %>%
  select(dPR1, dPR2, dPR3)

cor_diff <- cor(data_diff, use = "complete.obs")
round(cor_diff, 3)
```

```
##      dPR1 dPR2 dPR3
## dPR1 1.000 0.468 0.646
## dPR2 0.468 1.000 0.869
## dPR3 0.646 0.869 1.000
```

Correlation in first differences indicates shared volatility patterns, particularly between PR2 and PR3. However, this does not imply predictive causality, reinforcing the decision to avoid multivariate models such as VAR or ARIMAX.

The exploratory analysis reveals a demand process characterized by high volatility, abrupt shocks, and non-Gaussian changes, with limited data availability and no exogenous drivers. These findings motivate the use of robust, parsimonious models, the incorporation of recent shock features, and walk-forward validation to ensure stable out-of-sample performance.

## Modelling

### Techniques

Three approaches were evaluated using out-of-sample and walk-forward validation.

#### ARIMA (Benchmark)

Used as a standard univariate baseline. While it captured overall level and autocorrelation, ARIMA consistently smoothed abrupt demand shocks, leading to higher errors during volatile periods.

#### ETS (Discarded)

Smoothing-based models underperformed relative to ARIMA and failed to adapt to irregular demand patterns. The absence of stable seasonality made these models unsuitable for this task.

#### LightGBM (Final Model)

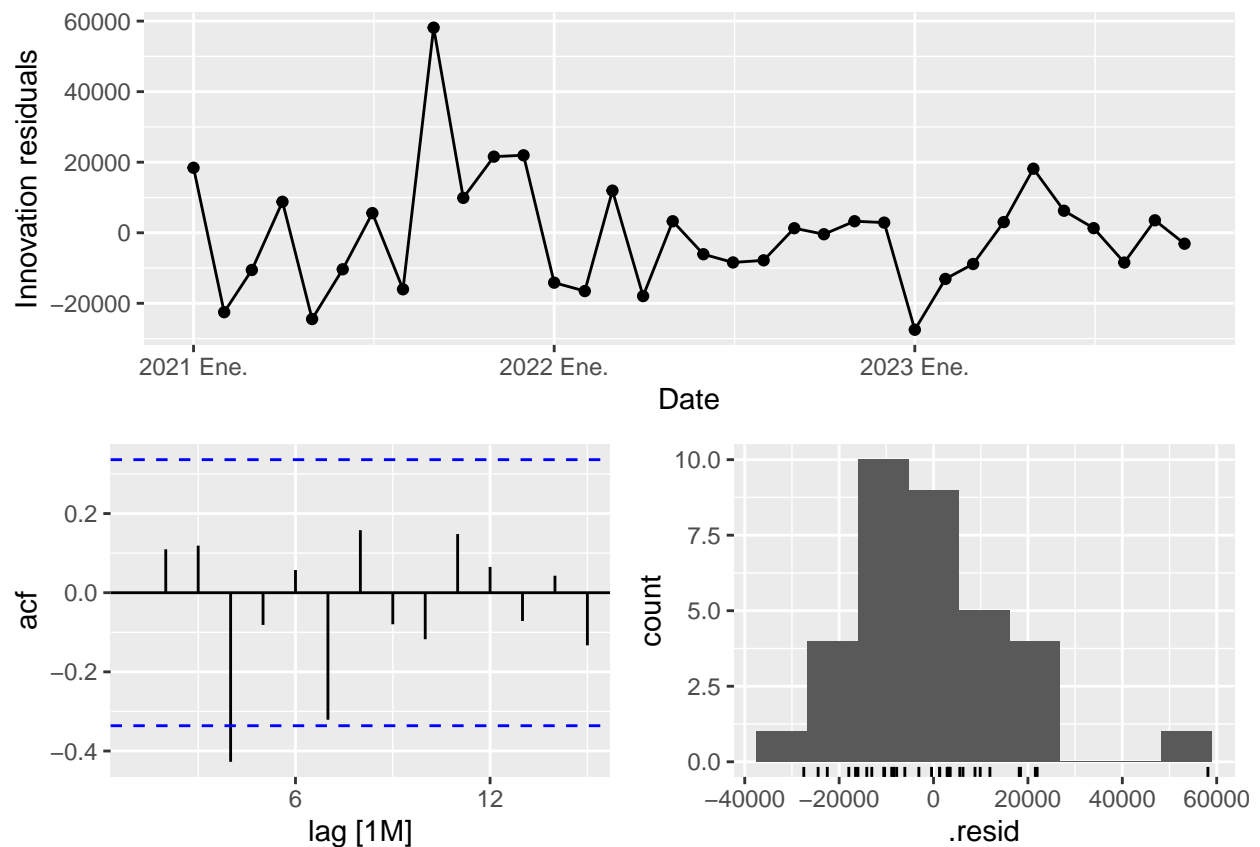
A feature-based LightGBM model outperformed all classical approaches. After incorporating recent shock features, it achieved a substantial reduction in forecast error and showed greater robustness during high-volatility periods.

## Modeling assumptions

- No exogenous variables available → feature-based approach required
- Limited sample size (40 obs) → parsimonious models, expanding window CV
- Irregular shocks → tree-based model with shock features
- Forecast horizon short (1–6 months) → no long-term trend modeling

```
#TRAIN/TEST
data_ts <- data_prd
train <- data_ts %>% filter(Date <= yearmonth("2023-10"))
test  <- data_ts %>% filter(Date >  yearmonth("2023-10"))
```

## ARIMA



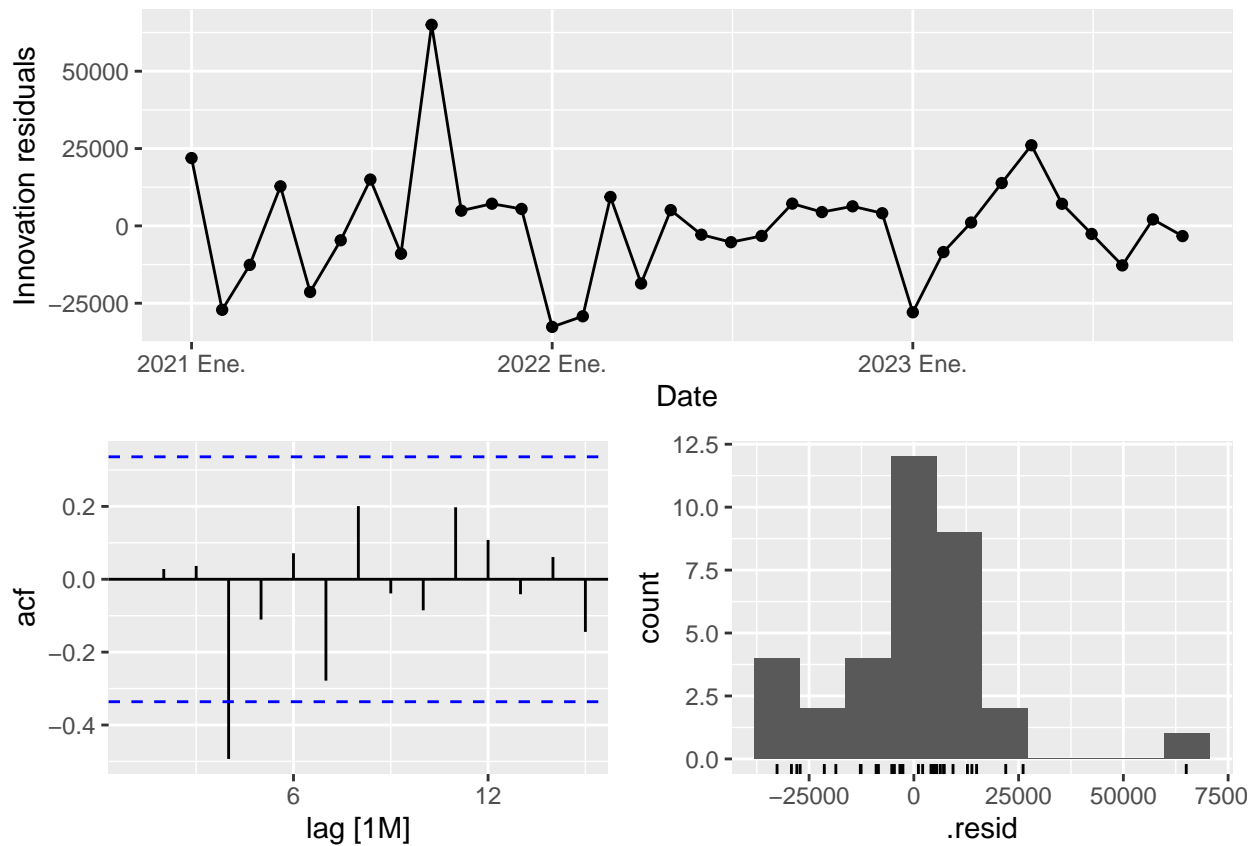
```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>   <dbl>   <dbl>
## 1 ARIMA    17.2    0.144
```

```
## # A tibble: 6 x 3
##   Date actual  pred
##   <mth> <dbl> <dbl>
## 1 2023 Nov.  72806 42263.
```

```
## 2 2023 Dic. 53050 42804.
## 3 2024 Ene. 46783 43114.
## 4 2024 Feb. 32730 43291.
## 5 2024 Mar. 69000 43393.
## 6 2024 Abr. 69273 43451.
```

```
##          RMSE          MAE          MAPE
## 2.035259e+04 1.774148e+04 2.929395e-01
```

## ETS



```
##          RMSE          MAE          MAPE
## 2.097724e+04 1.831821e+04 2.997203e-01
```

## LGBM

Classical time series models struggled to adapt to abrupt demand shocks and irregular volatility. To address these limitations, a LightGBM regression model was evaluated as a flexible, non-linear alternative capable of incorporating engineered temporal features and learning complex patterns from limited data.

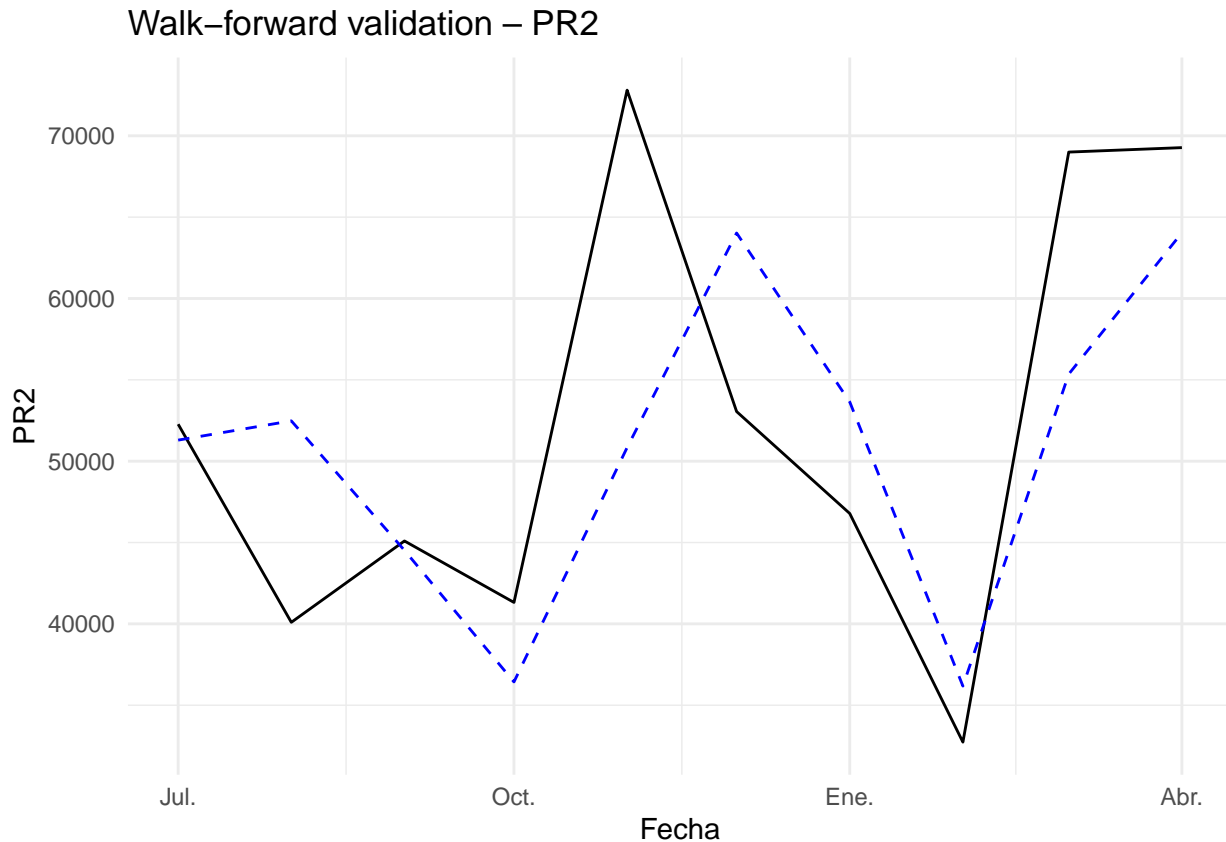
Features included

- Lagged demand (1, 2, 3, 6, 12 months)
- Rolling averages (3, 6, 12 months) to capture local context

- Cyclical seasonality using sine and cosine transformations

Model performance was evaluated using walk-forward (expanding window) validation, retraining the model at each step and forecasting one month ahead. This approach reflects real-world deployment conditions and prevents look-ahead bias.

```
##          RMSE          MAE          MAPE
## 1.026653e+04 8.094788e+03 1.492352e-01
```



Model	RMSE	MAPE
ARIMA (benchmark)	20000	29%
ETS	21000	30%
LightGBM (basic)	17000	26%
<b>LightGBM + shock features</b>	<b>10000</b>	<b>15%</b>

Reducing MAPE from aprox. 30% to aprox. 15% implies halving the average relative forecast error, which translates into materially lower safety stock requirements under volatile demand.

## Scenarios

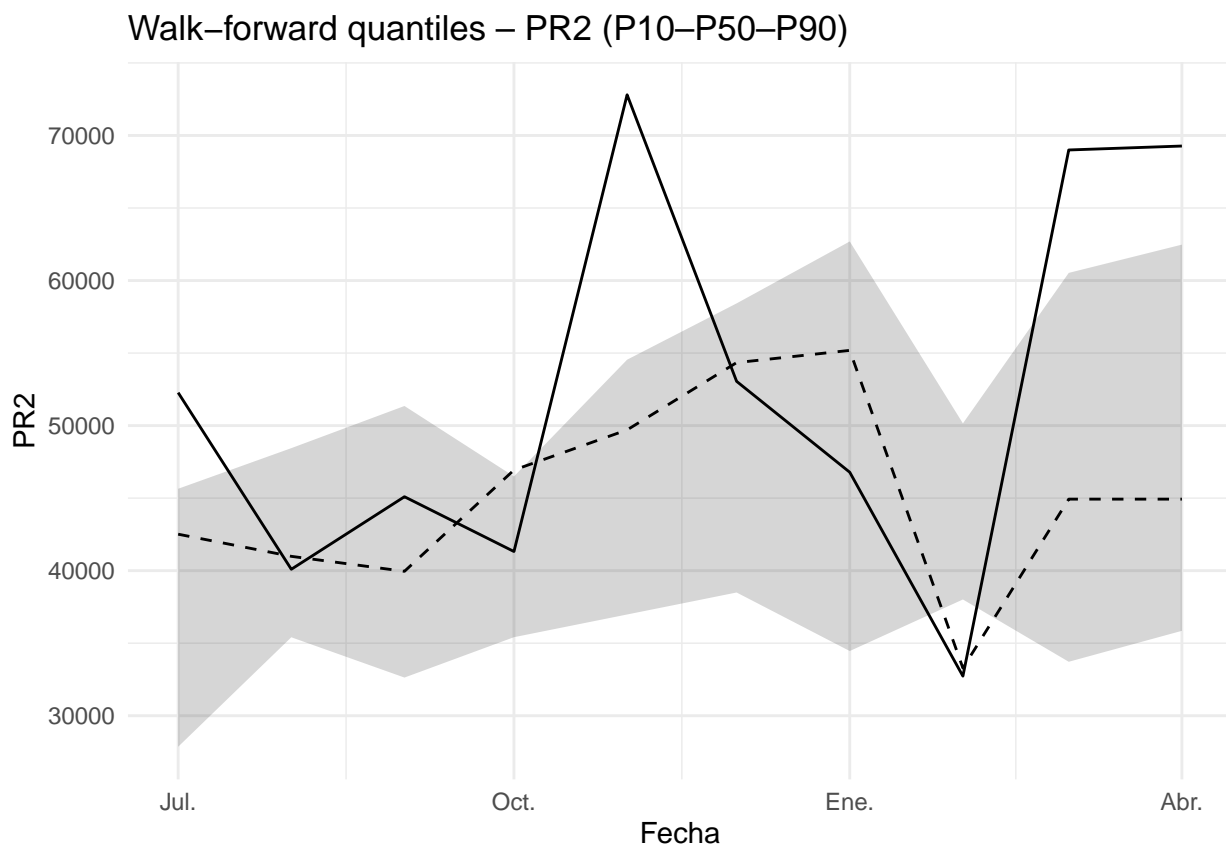
P10 represents a conservative scenario for downside planning, while P90 captures potential demand surges. This enables inventory policies to be stress-tested under uncertainty rather than relying on a single point estimate.



```
## # A tibble: 6 x 5
##   date          actual    p10    p50    p90
##   <dtm>         <dbl>  <dbl>  <dbl>  <dbl>
## 1 2023-07-01 00:00:00 52273 27860. 42513. 45645.
## 2 2023-08-01 00:00:00 40096 35409. 40987. 48445.
## 3 2023-09-01 00:00:00 45095 32630. 39953. 51352.
## 4 2023-10-01 00:00:00 41319 35413. 46936. 46495.
## 5 2023-11-01 00:00:00 72806 36978. 49705. 54543.
## 6 2023-12-01 00:00:00 53050 38495. 54340. 58424.
```

```
## [1] 0.5
```

```
## [1] 19192.64
```



## Final Forecasts

Final forecasts were generated using the LightGBM model with shock-aware features, selected based on superior walk-forward performance and robustness under demand volatility. Forecasts are produced for the next six months and reported using quantile estimates (P10–P50–P90) to reflect uncertainty and support scenario-based planning.

## Forecasts

tbl\_PR1

```
## # A tibble: 6 x 4
##   Month      P10    P50    P90
##   <chr>    <dbl> <dbl> <dbl>
## 1 2024-05 4904. 7578. 7705.
## 2 2024-06 6912. 8204. 8897.
## 3 2024-07 6164. 7793. 7592.
## 4 2024-08 7729. 8362. 8423.
## 5 2024-09 7755. 7972. 7785.
## 6 2024-10 7758. 8963. 9210.
```

tbl\_PR2

```
## # A tibble: 6 x 4
##   Month      P10    P50    P90
##   <chr>    <dbl> <dbl> <dbl>
## 1 2024-05 36535. 57220. 64058.
## 2 2024-06 37102. 46579. 50952.
## 3 2024-07 37017. 43890. 50950.
## 4 2024-08 35586. 43195. 45756.
## 5 2024-09 36848. 41981. 44281.
## 6 2024-10 38099. 46483. 41918.
```

tbl\_PR3

```
## # A tibble: 6 x 4
##   Month      P10    P50    P90
##   <chr>    <dbl> <dbl> <dbl>
## 1 2024-05 48236. 61138. 64002.
## 2 2024-06 46926. 55709. 66943.
## 3 2024-07 42170. 58844. 74349.
## 4 2024-08 45105. 55679. 71008.
## 5 2024-09 46454. 57483. 69251.
## 6 2024-10 46532. 57862. 60544.
```