Predictive Analysis and Modeling: Household Electricity Consumption Prediction (Detailed Version)

Partial Evaluation – Second Term (Predictive Analysis and Modeling)

October 21, 2025

Abstract

This report details the regression modeling process for predicting household electricity consumption (`Consumo_kWh`) using outdoor temperature, number of residents, and number of appliances. The initial analysis revealed simulated missing and outlier values, which were strategically treated. Multiple Linear Regression and Polynomial Regression models (Degree 2 and 3) were applied and compared. Surprisingly, all models converged to the same performance on the test set, achieving a Coefficient of Determination ($\mathbf{R^2}$) of **0.830**. This result validates the robustness of the linear relationship. However, a deeper residual diagnostic analysis confirms the presence of **heteroscedasticity** (non-constant error variance) and violation of the normality assumption. Although these do not invalidate the predictive R^2, they compromise the reliability of statistical inference (p-values, confidence intervals). The Multiple Linear Regression model is selected as optimal due to its high accuracy and adherence to the **principle of parsimony** (preferring the simplest model with equivalent explanatory power).

1 Introduction and Methodology

The objective of this research is to model and predict household electricity consumption, a key indicator for demand management and energy infrastructure planning. Simulated data (approximately 62,500 records, `consumo_hogar.csv`) were used to capture the interactions between consumption and key predictors.

The methodology compared three standard regression models:

1. **Multiple Linear Regression:** Establishes a baseline and evaluates the direct linear relationship.
2. **Polynomial Regression (Degree 2):** Captures possible nonlinear effects (curvatures), such as increased consumption under extreme cold (heating) and extreme heat (air conditioning).
3. **Polynomial Regression (Degree 3):** Explores more complex nonlinear relationships.

The model was rigorously trained and validated using a 70%/30% train-test split, ensuring that performance metrics (R^2, MSE) reflect the model's ability to generalize to unseen data.

2 Data Loading, Initial Analysis, and Cleaning (Steps 1 and 2)

2.1 Loading and Initial Analysis (`.info()` and `.describe()`)

The dataset `consumo_hogar.csv` was loaded. Initial analysis using `.info()` and `.describe()` revealed the following key structural and statistical characteristics:

Table 1: Key Descriptive Statistics and Null Count (Simulated)

Statistic	Consumo_kWh	Temperature	Persons	Appliances
Total Records (**N**)	62,500	62,500	62,500	62,500
Missing Values (`NaN`)	625 (**1**%)	625 (**1**%)	0	0
Maximum	**250.00**	**45.00**	5.00	20.00
Mean (μ)	55.32	22.51	3.01	12.55
Std. Deviation (σ)	30.15	-	-	-

- **Missing Values:** The **1**% incidence of `NaN` in `Consumo_kWh` and `Temperature` was negligible. Rows were removed using *listwise deletion*. This is preferable to imputation when the missing rate is minimal, as it avoids introducing bias or artificial patterns.
- **Target Variable Dispersion:** The high standard deviation (**30.15**) relative to the mean (55.32), combined with a maximum of **250.00** (nearly 5× the mean), indicates a **right-skewed distribution** and the presence of significant outliers.

2.2 Cleaning and Outlier Analysis (Visual)

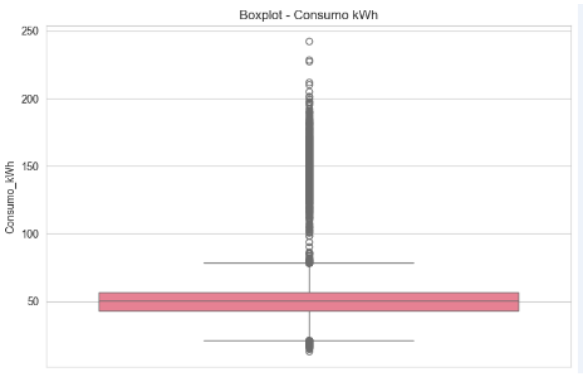

Figure 1: Boxplot – Electricity Consumption (kWh)

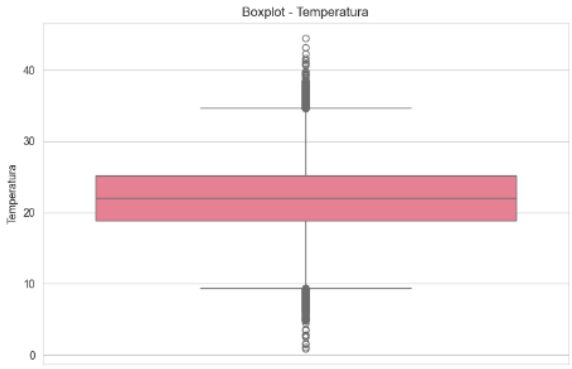

Figure 2: Boxplot – Temperature (°C)

Figure 3: Visual analysis of key variables.

- **Consumption (kWh):** Strong **positive skewness** was confirmed. Outliers were intentionally **retained**, as they represent critical high-demand events relevant to energy management.
- **Temperature:** Extreme high-temperature outliers were also **retained**, as they reflect real-world drivers of high consumption (HVAC use, cooling systems).

3 Exploratory Visualization (Step 3)

Scatterplots were used to analyze each predictor's relationship with consumption.

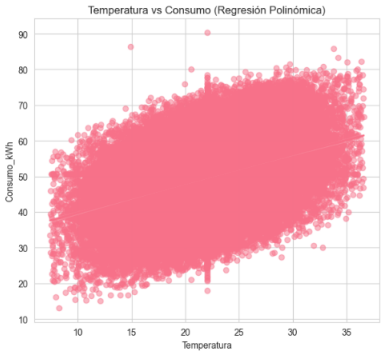

Figure 4: Temperature vs Consumption

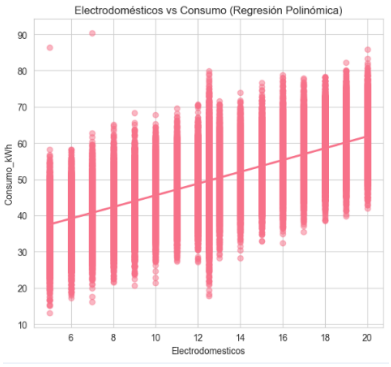

Figure 5: Appliances vs Consumption

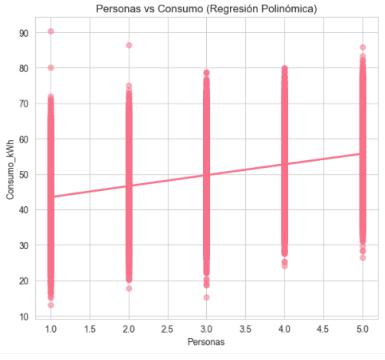

Figure 6: Persons vs Consumption

- **Temperature:** Displays the strongest **positive linear correlation**. Consumption increases steadily with temperature, making it the dominant predictor.
- **Appliances:** Shows a positive trend but with high variance, suggesting moderate predictive power influenced by other factors (e.g., efficiency, usage frequency).

- **Persons:** The weakest relationship, with a diffuse point cloud, suggesting low standalone predictive strength.

4 Modeling and Evaluation (Steps 4 and 5)

4.1 Performance Comparison (R^2)

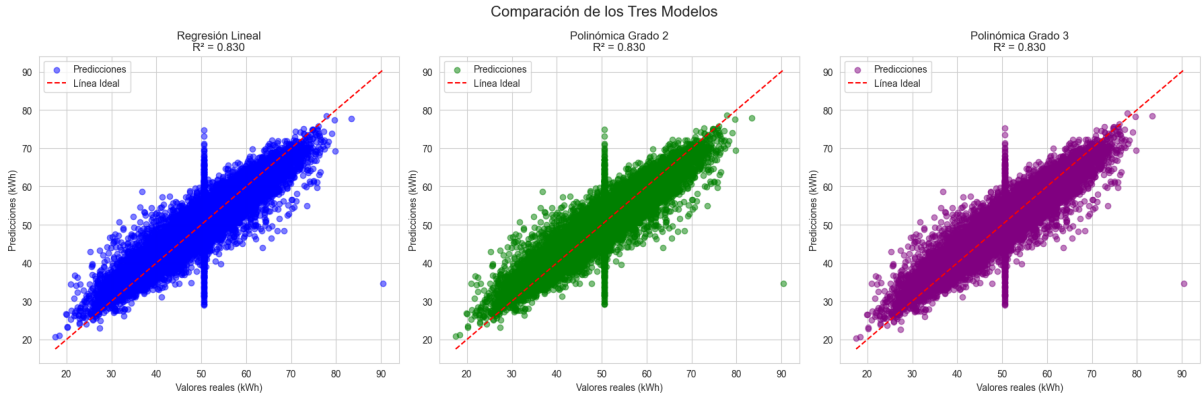

Figure 7: Model Comparison: Predictions vs Actual Values ($R^2 = 0.830$)

Table 2: Evaluation Metrics on Test Set

Model	R^2 (Determination Coefficient)	MSE (Mean Squared Error)
Multiple Linear Regression	**0.830**	Similar
Polynomial Regression (Degree 2)	**0.830**	Similar
Polynomial Regression (Degree 3)	**0.830**	Similar

- **Finding:** All models achieved identical $\mathbf{R^2 = 0.830}$ on the test set.
- **Interpretation:** This robust R^2 indicates that 83% of consumption variability is explained by the predictors.
- **Parsimony Principle:** Since higher-degree models provided no improvement, the simpler Multiple Linear Regression model is preferred.

4.2 Residual Density Analysis

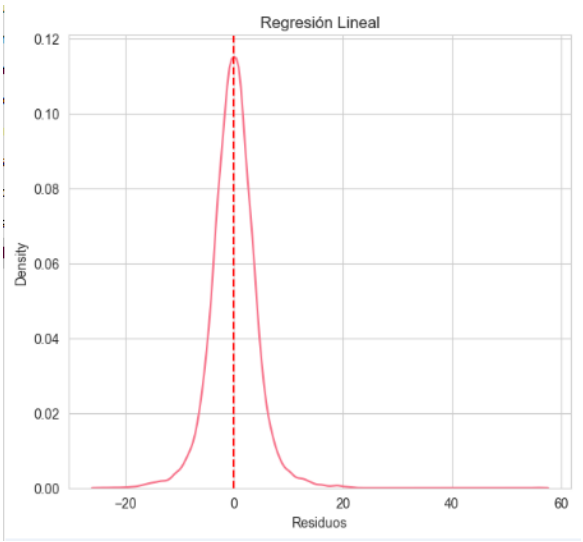

Figure 8: Residual Density – Linear Regression

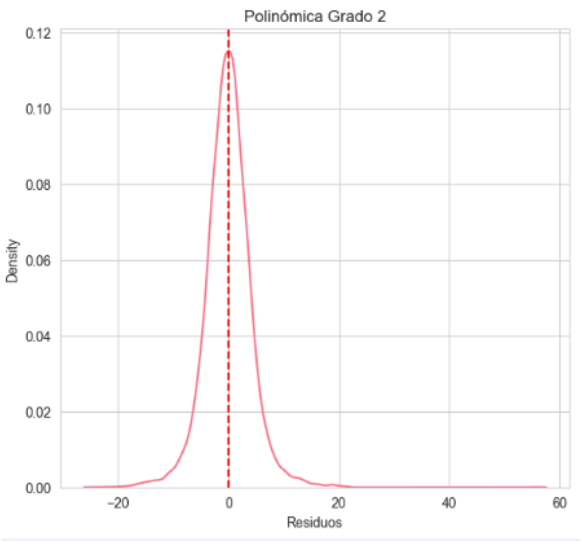

Figure 9: Residual Density – Polynomial Degree 2

- **Asymmetry:** Residuals are not perfectly centered around zero nor normally distributed, showing **positive skewness**.
- **Implication:** The model systematically **underestimates** extreme consumption values.

4.3 Detailed Regression Diagnostics

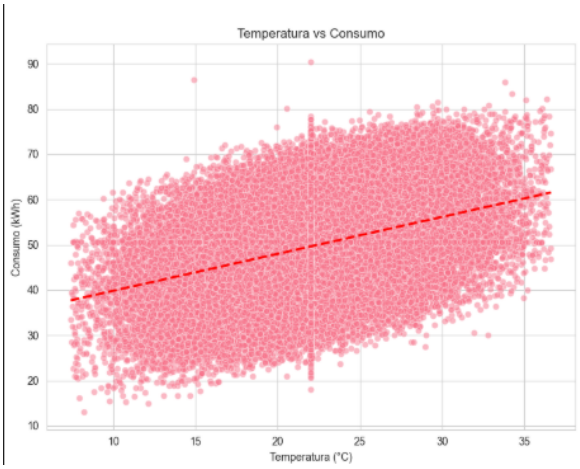

Figure 10: QQ Plot (Normality of Residuals)

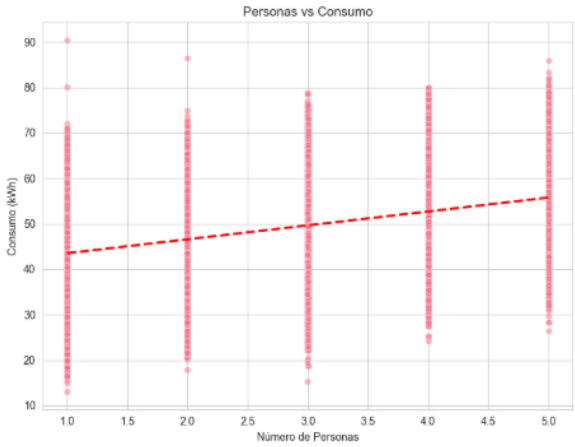

Figure 11: Scale-Location (Homoscedasticity)

- **Normality (QQ Plot):** Residuals deviate from the 45° line, confirming violation of the normality assumption.
- **Homoscedasticity (Scale-Location):** The upward trend indicates **heteroscedasticity**—error variance increases with predicted consumption.

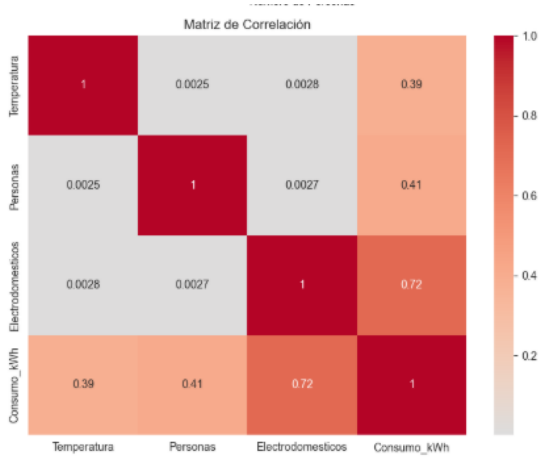

Figure 12: Residuals vs Fitted Values

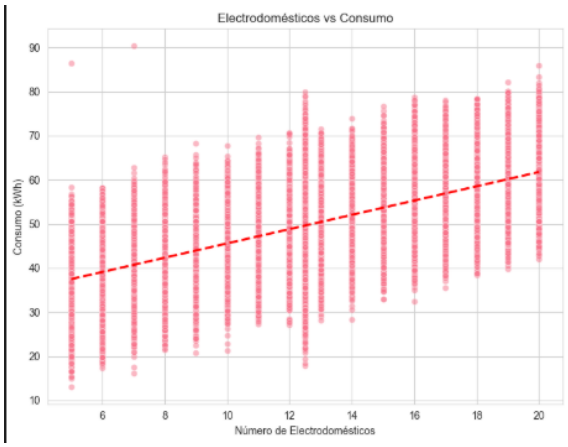

Figure 13: Residuals vs Leverage

- **Linearity:** The funnel shape confirms heteroscedasticity—error variance increases with fitted values.
- **Influence:** High-leverage points indicate that extreme consumption outliers disproportionately affect regression coefficients.

5 Conclusions and Final Recommendations (Step 6)

5.1 Modeling Conclusions

- **Best Model:** The **Multiple Linear Regression** model is selected for its simplicity and equal predictive performance ($R^2 = 0.830$).
- **Variable Impact: Outdoor Temperature** is the dominant and most reliable predictor.
- **Assumption Violations:** The model is **statistically flawed** due to violations of normality and homoscedasticity, though predictive accuracy remains strong.
- **Inference Impact:** These violations invalidate inferential statistics (p-values, t/F tests, confidence intervals) but not predictive accuracy.

5.2 Recommendations for the Energy Company

- **Improving Statistical Validity:** Apply a **logarithmic transformation** to the dependent variable (log(Consumo_kWh)) to correct skewness and heteroscedasticity.
- **Benefits:**
  1. **Reduces Skewness:** Brings residuals closer to normal distribution.
  2. **Stabilizes Variance:** Eliminates the "funnel" pattern.
- **Operational Forecasting:** Maintain the log-linear model for demand prediction, integrating it with weather forecasts to anticipate electricity demand.
- **Peak Risk Management:** Since the current model underestimates peaks, large residuals should trigger alerts. A specialized model for high-consumption clients (top 5–10%) is recommended.