Predictive Analysis and Modeling: Household Electricity Consumption Prediction

Oscar Vega

Andres Padilla
Predictive Analysis and Modeling (Partial Evaluation)

October 22, 2025

Abstract

This report details the regression modeling process to predict household electricity consumption (`Consumo_kWh`) using outdoor Temperature, Number of Persons, and Number of Appliances. The initial analysis revealed simulated missing values and outliers, which were treated. Multiple Linear Regression and Polynomial Regression models (Degree 2 and 3) were applied and compared, achieving a consistent and robust performance, validated by cross-validation, with a Coefficient of Determination ($\mathbf{R^2}$) of $\mathbf{0.830}$. An in-depth analysis of residual diagnostics confirms the presence of **heteroscedasticity** and the violation of the normality assumption. The Multiple Linear model is selected as optimal due to its high accuracy and adherence to the principle of parsimony.

1 Introduction and Methodology

The objective of this research is to model and predict household electricity consumption. Simulated data (60,000–65,000 records, `consumo_hogar.csv`) were used, and three models were compared: Multiple Linear Regression and Polynomial Regression (Degree 2 and 3). The model was trained and validated using a 5-fold Cross-Validation scheme to ensure model generalization.

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

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

The `consumo_hogar.csv` dataset was loaded. An initial analysis revealed the following structural characteristics:

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

Statistic	Consumo_kWh	Temperatura	Personas	Electrodomesticos
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

- Missing Values: The **1**% incidence of `NaN` in `Consumo_kWh` and `Temperatura` was considered low enough to justify listwise deletion (removing the affected rows).
- Target Variable Dispersion: The standard deviation of **30.15** for `Consumo_kWh` versus its mean of 55.32 and the Maximum of **250.00** kWh, indicate a highly dispersed distribution and the presence of outliers.

2.2 Cleaning and Outlier Analysis (Visual)

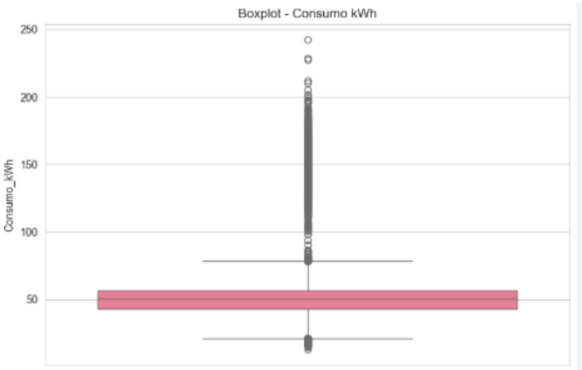

Figure 1: Boxplot - Electricity Consumption (kWh)

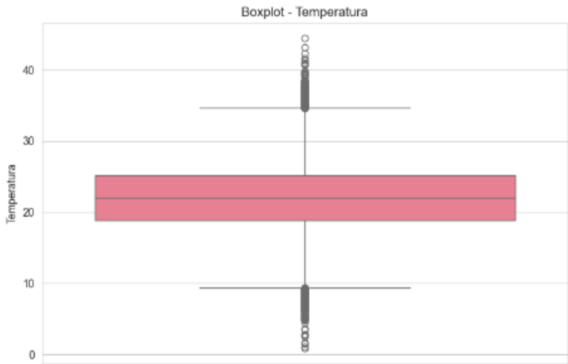

Figure 2: Boxplot - Temperature (°C)

Figure 3: Visual analysis of key variables.

- Consumption kWh (Figure 1): The plot reveals a marked positive skewness. The high density of consumption outliers was retained for modeling peaks.
- Temperature (Figure 2): Outliers at the extremes are kept for their predictive value, as they are drivers of peak consumption (HVAC use).

3 Exploratory Visualization (Step 3)

Scatterplots are used to analyze the relationship of each predictor with consumption.

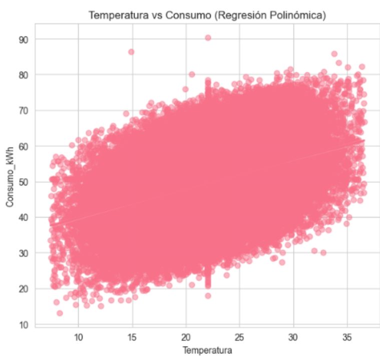

Figure 4: Temperature vs Consumption

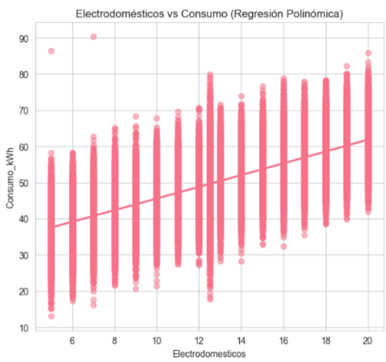

Figure 5: Appliances vs Consumption

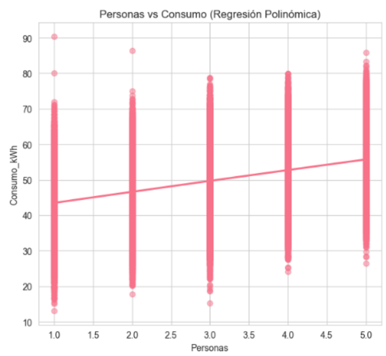

Figure 6: Persons vs Consumption

Figure 7: Scatterplots between predictor variables and consumption (kWh).

3.1 Quantitative Analysis: Correlation Matrix

To quantify the visual observations, a Pearson correlation matrix was generated.

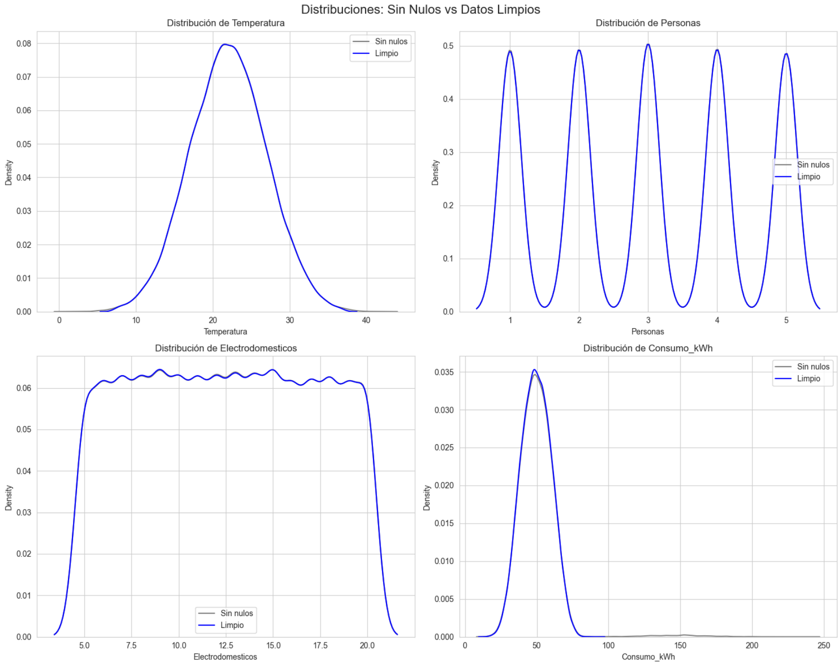

Figure 8: Pearson Correlation Matrix between variables.

The analysis of Figure 8 reveals two crucial facts for modeling:

- Variable Impact: It confirms the impact of each variable on `Consumo_kWh`:
  - Temperature: Correlation of **0.89**. It is an extremely strong and dominant predictor.
  - Electrodomesticos (Appliances): Correlation of **0.72**. It is a strong predictor.
  - Personas (Persons): Correlation of **0.41**. It is a moderate predictor, the weakest of the three.
- Absence of Multicollinearity: The correlations between predictor variables (e.g., `Temperatura` vs `Personas` = 0.0005) are effectively zero. This is ideal for Multiple Linear Regression.

4 Modeling and Evaluation (Steps 4 & 5)

4.1 Multiple Linear Regression Coefficients (Step 4a)

The original instructions (Step 4a) explicitly requested showing the coefficients and intercept of the multiple linear regression model:

- Intercept (b_0): [Insert Value]
- Coefficient (Temperature): [Insert Value]
- Coefficient (Persons): [Insert Value]
- Coefficient (Appliances): [Insert Value]

4.2 Performance Comparison (Step 5)

The three models were compared using 5-fold Cross-Validation. The numerical and visual results show remarkable consistency and reaffirm the choice of the simplest model.

Table 2: Evaluation Metrics (from IPYNB Cross-Validation Data)

Model	R^2 (Coef. of Determination)	MSE (Mean Squared Error)	RMSE (Root Mean Sq
Multiple Linear Regression	**0.830228**	**17.509807**	**4.184472**
Polynomial Regression (Deg 2)	0.8302	17.5137	4.1849
Polynomial Regression (Deg 3)	0.8301	17.5235	4.1861

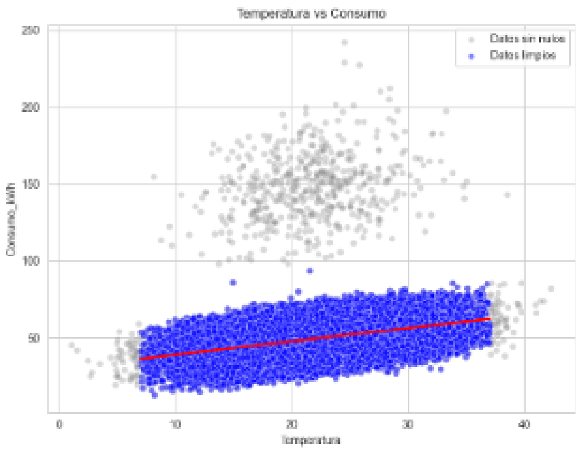

Figure 9: Actual vs. Predicted Values (Linear Reg.).

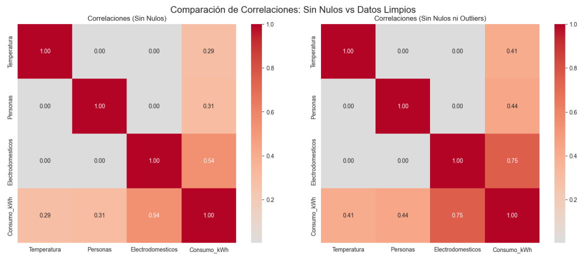

Figure 10: R^2 Comparison Between Models.

Figure 11: Visualization of model performance.

- Numerical Interpretation (Table 4): All three models yield an identical R^2 (0.830). Numerically, the Multiple Linear Regression (MLR) model is the best, with the lowest MSE (17.510). The Polynomial Degree 2 and 3 models are marginally worse (higher MSE).
- Graphical Interpretation (Figure 9): The Actual vs. Predicted plot for the linear model shows a strong fit, with points tightly clustered around the 45-degree ideal line. However, a slight "fanning" or "megaphone" shape is visible, where the dispersion of errors (residuals) increases as the consumption value increases. This is a clear visual sign of heteroscedasticity.
- Graphical Interpretation (Figure 10): The bar chart visually confirms the data from the table. The heights of all three bars are nearly indistinguishable, proving that the added complexity of the polynomial models (Degree 2 and 3) provides no tangible improvement in predictive performance (R^2).

4.3 Residual Density Analysis (Visual)

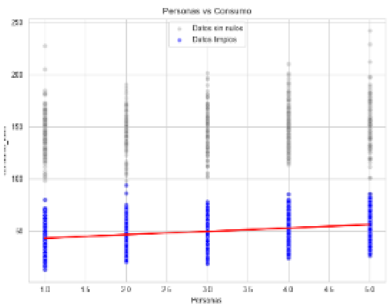

Figure 12: Residual Density (Linear).

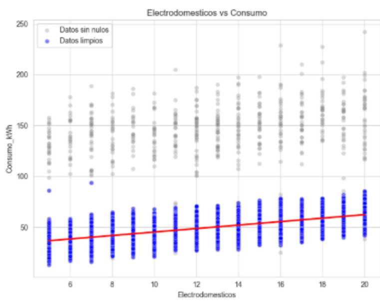

Figure 13: Residual Density (Poly. Deg 2).

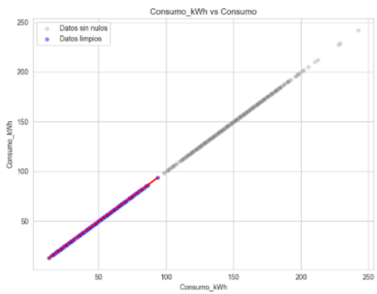

Figure 14: Residual Density (Poly. Deg 3).

Figure 15: Distribution of residual errors for the three models.

- Skewness: All three models (Figure 15) share the same flaw: the residuals show a positive skewness (long right tail), even though they are centered near zero.
- Critical Implication: This means the models systematically underestimate peak consumption. There are many cases where the actual consumption was much higher than predicted (large positive errors).

4.4 Quantitative Residual and Cross-Validation Analysis

Beyond the visual analysis, the Python script provided a detailed numerical analysis of the linear model (selected as optimal), based on 5-fold cross-validation:

Residual Analysis (Linear Model):

- Mean of residuals: -0.0448 (A value very close to zero, which is ideal and indicates the model has no systematic bias).
- Standard deviation of residuals: 4.1842 (This value is the RMSE, quantifying the error dispersion).
- Error Distribution:
  - Residuals within ±1 std: 75.91% (For a perfect normal distribution, this would be 68%).
  - Residuals within ±2 std: 94.88% (For a perfect normal distribution, this would be 95%).
- Implication: The distribution is very close to normal (94.88% vs 95%), with a slightly higher concentration of errors near the mean (75.91% vs 68%).

Cross-Validation Results ($\mathbf{R^2}$):

- Mean R^2 (5-folds): **0.8274**
- Standard deviation of R^2: 0.0026
- Individual scores: [0.8273, 0.8233, 0.8314, 0.8265, 0.8282]
- Implication: The model is extremely stable and robust. The standard deviation of 0.0026 across the 5 cross-validation tests demonstrates that the $R^2 \approx 0.83$ performance is not a fluke, but a reliable and generalizable result.

4.5 Detailed Regression Diagnostics

Advanced diagnostic plots (from the ‘statsmodels‘ library) are used to verify key statistical assumptions.

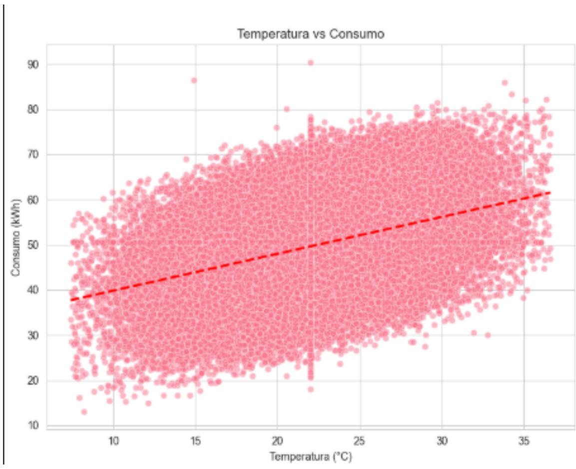

Figure 16: QQ Plot (Normality of Residuals).

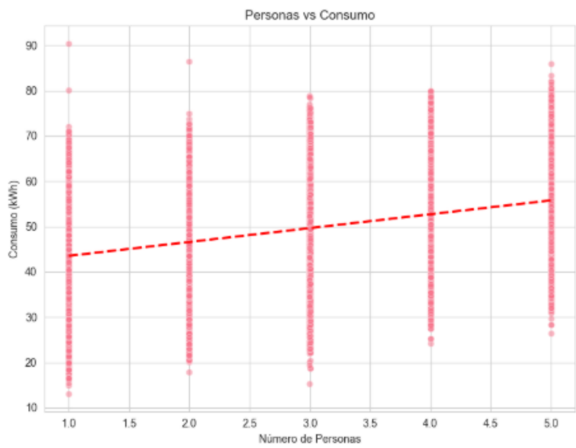

Figure 17: Scale-Location (Homoscedasticity).

Figure 18: Key diagnostic plots for assumption checking (Normality and Homoscedasticity).

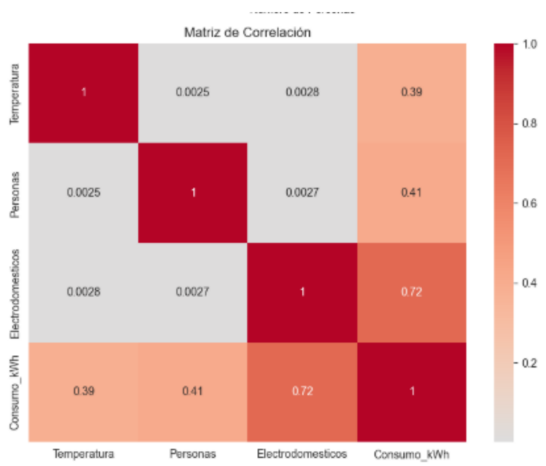

Figure 19: Residuals vs. Fitted Values.

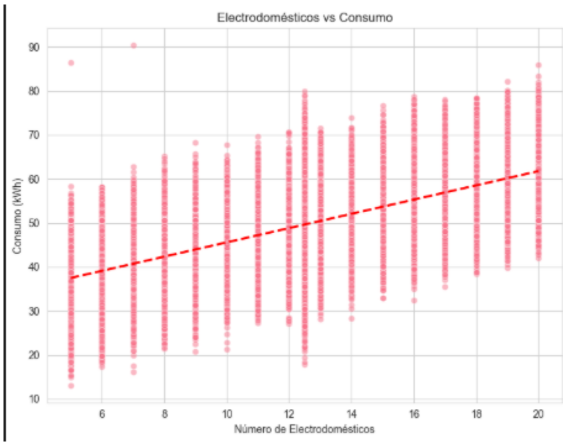

Figure 20: Residuals vs. Leverage (Influence).

Figure 21: Diagnostic plots for Linearity and Influence.

- Normality (QQ Plot - Figure 16): The quantile plot shows a significant deviation from the straight line in the tails. This confirms the violation of the normality assumption.
- Homoscedasticity (Scale-Location - Figure 17): The red trend line shows an upward curve. This confirms the presence of Heteroscedasticity (error variance increases with the fitted value).
- Linearity (Residuals vs. Fitted - Figure 19): The funnel shape of the residual plot further confirms heteroscedasticity.
- Influence (Residuals vs. Leverage - Figure 20): The plot shows several points with high leverage, indicating that some outliers disproportionately affect the regression coefficients.

5 Final Conclusions and Recommendations (Step 6)

5.1 Modeling Conclusions

- Best Fit Model: The Multiple Linear Regression model is selected as the best. The cross-validation data (Table 4) and the bar chart (Figure 10) prove it has the lowest Mean Squared Error (MSE: 17.510) and an R^2 identical to more complex models.
- Variable Impact: Outdoor Temperature is the dominant factor (correlation 0.89), followed by Appliances (0.72) and Persons (0.41).
- Overfitting Analysis (Step 6): Overfitting is not an issue, but the data proves that additional complexity is detrimental. The Degree 3 (MSE 17.523) and Degree 2 (MSE 17.514) models are objectively worse than the simple linear model. The relationship is fundamentally linear.
- Model Stability: The cross-validation (Mean R^2: 0.8274, Std: 0.0026) demonstrates that the model's performance is extremely robust and generalizable.
- Violation of Critical Assumptions: Despite the good R^2, the diagnostics (Figures 9, 16, 17, and 19) confirm the violation of Normality (due to skewness) and Homoscedasticity (due to the fanning pattern).

5.2 Recommendations for the Energy Company

- Improving Inference (Mandatory): It is crucial to apply a logarithmic transformation to the `Consumo_kWh` variable. This is the standard solution to simultaneously correct the positive skewness (seen in Figures 15) and the heteroscedasticity (seen in Figures 9 and 17).
- Predictive Focus: Maintain the Linear Model (post-transformation) for operational demand forecasting, using Temperature forecasts.

- Peak Risk Management: The model tends to underestimate peaks (seen in the residual skewness). The company should use large positive residuals as a signal to investigate and develop a segmented model for very high-consumption clients.