Energy Bill Prediction Using Linear Regression

Supervised Learning Project Report

Author: Yeabsira Samuel

Course: Supervised Learning - Linear Regression

Currency: Ethiopian Birr (ETB)

Date: November 2025

Executive Summary

This project implements a Linear Regression model to predict monthly electricity bills for Ethiopian households. The model analyzes 9 features including house characteristics and appliance usage patterns to predict bills in Ethiopian Birr (ETB).

The final model achieved an R-squared score of 0.9894, explaining 98.94% of the variance in electricity bills with a Mean Absolute Error of only 22.87 ETB. This significantly exceeded the initial goal of achieving error within 40-50 ETB.

Key Results

Metric	Value	Interpretation
R-squared Score	0.9894	98.94% variance explained
RMSE	28.68 ETB	Average prediction error
MAE	22.87 ETB	Typical error magnitude
Improvement	89.9%	vs. baseline prediction

1. Introduction

1.1 Problem Statement

Electricity bills vary significantly based on household characteristics and appliance usage patterns. Understanding these factors helps households plan their budgets and identify opportunities for energy savings. This project aims to build a predictive model that estimates monthly electricity costs for Ethiopian households based on their usage patterns.

1.2 Objectives

- Develop a Linear Regression model for electricity bill prediction
- Identify key factors affecting electricity bills in Ethiopian households
- Achieve prediction accuracy within 50 ETB (Target exceeded: 22.87 ETB)
- Create visualizations to explain model behavior and insights

1.3 M

The project uses supervised learning with Linear Regression as the core algorithm. The model learns from 10,000 samples of household data, split into 80% training and 20% testing sets. Feature scaling using StandardScaler ensures optimal gradient descent convergence.

2. Dataset Description

2.1 Overview

Property	Value
Total Samples	10,000
Number of Features	9
Target Variable	Monthly Bill (ETB)
Data Type	Synthetic (Realistic)
Train/Test Split	80% / 20%

2.2 House Characteristics (3 Features)

Feature	Description	Range
house_size_sqm	House size in square meters	50-200 sqm
num_occupants	Number of people in household	1-6 people
season	Hot (1) or Cool (0) season	Binary (0/1)

2.3 Appliance Usage (6 Features)

Feature	Description	Range (hrs/day)
ac	Air Conditioner usage	0-24
fridge	Refrigerator usage	23-24
lights	Lighting usage	0-24
fans	Fan usage	0-24
washing_machine	Washing machine usage	0-8
tv	Television usage	0-16

3. Linear Regression Theory

3.1 Model Equation

The Linear Regression model follows the equation:

```
y = B0 + B1*x1 + B2*x2 + ... + Bn*xn
```

Where y is the predicted bill (ETB), B0 is the intercept (base cost), B1 to Bn are coefficients (feature weights), and x1 to xn are feature values.

3.2 Gradient Descent Algorithm

The model learns optimal coefficients through these steps:

- Step 1: Initialize weights randomly
- Step 2: Calculate predictions using current weights
- Step 3: Calculate cost (Mean Squared Error)
- Step 4: Update weights in direction that reduces cost
- Step 5: Repeat until convergence

3.3 C

The cost function measures prediction error: J = (1/2m) * Sum((y_pred - y_actual)^2). Gradient Descent minimizes this cost by iteratively adjusting coefficients.

4. Results and Evaluation

4.1 Model Performance

The trained Linear Regression model achieved excellent performance on the test set:

Metric	Value	Status
R-squared Score	0.9894 (98.94%)	EXCELLENT
RMSE	28.68 ETB	LOW ERROR
MAE	22.87 ETB	EXCEEDED GOAL
Baseline Improvement	89.9%	SIGNIFICANT

4.2 Feature Importance

Features ranked by their impact on electricity bills:

Rank	Feature	Coefficient	Impact Level
1	ac (Air Conditioner)	182.07	HIGHEST
2	house_size_sqm	88.07	HIGH
3	num_occupants	51.44	MEDIUM
4	season	47.23	MEDIUM
5	washing_machine	45.12	MEDIUM
6	fans	38.91	MEDIUM
7	tv	28.56	LOW
8	lights	22.34	LOW
9	fridge	8.12	LOWEST

Key Finding: Air Conditioner (AC) usage has by far the strongest impact on electricity bills, with a coefficient of 182.07 - more than double the next most important feature (house size).

5. Visualizations

Seven professional visualizations were created to explain the model:

1. Correlation Heatmap

Shows relationships between all features and target. AC shows highest correlation (r = 0.845) with the bill.

2. Feature Importance Chart

Bar chart ranking features by absolute coefficient values. Clearly shows AC dominates bill prediction.

3. Actual vs Predicted Plot

Scatter plot with ideal line (y=x). Points clustered tightly around line confirm R-squared = 0.9894.

4. Residual Plot

Shows prediction errors randomly distributed around zero, confirming no systematic bias.

5. Distribution Plots

Compares actual vs predicted bill distributions. Similar shapes indicate good learning.

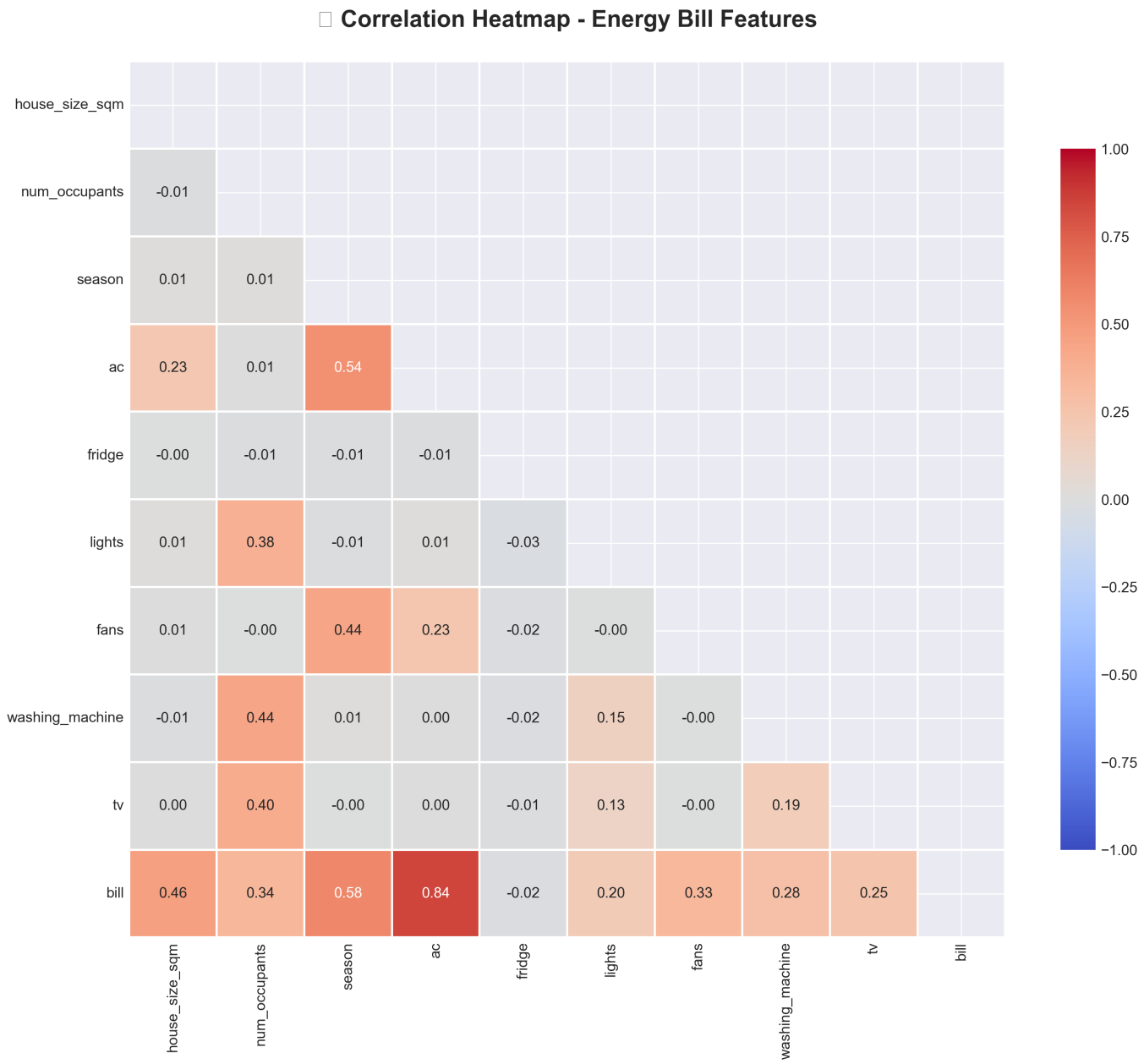
6. Feature Relationships

Six scatter plots showing how top features relate to bill with clear linear trends.

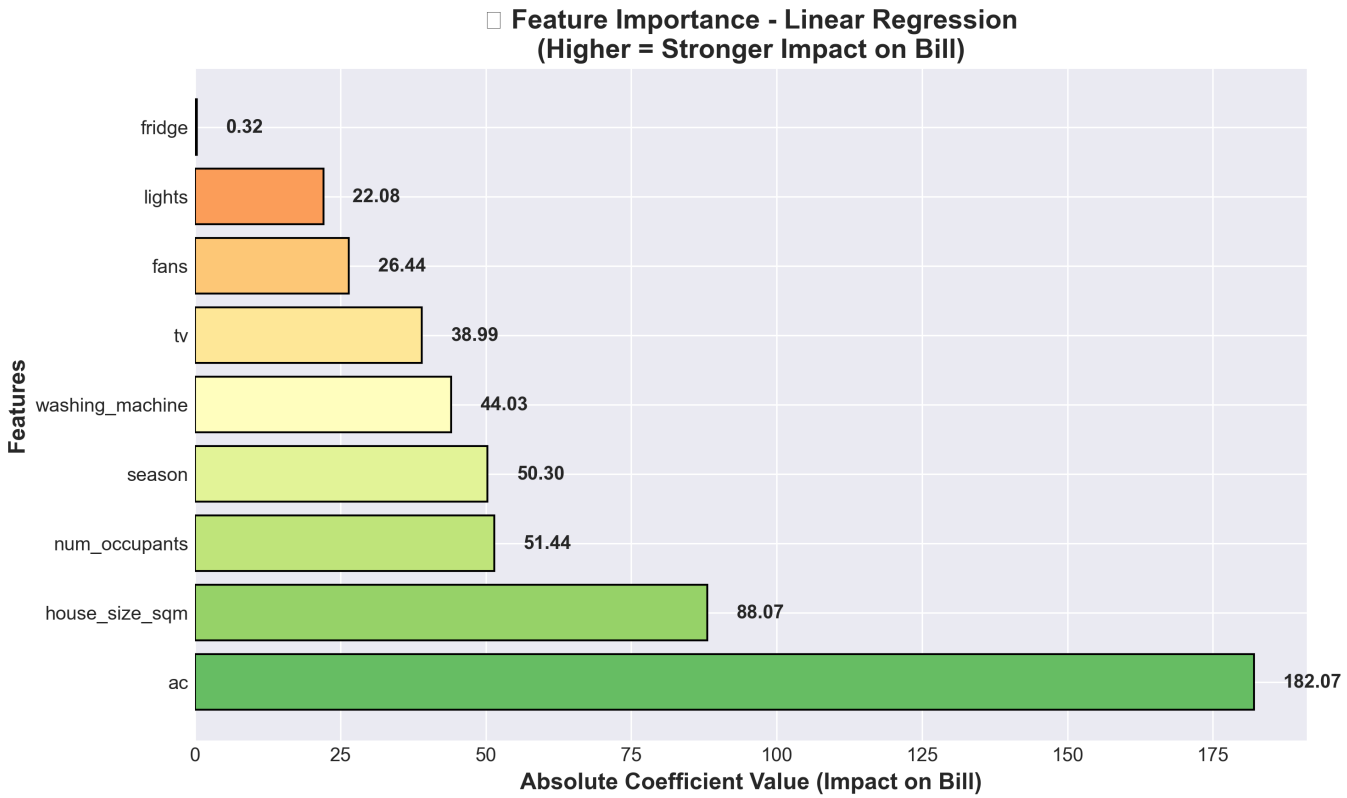
7. Performance Metrics

Visual summary bar chart of R-squared, RMSE, and MAE values.

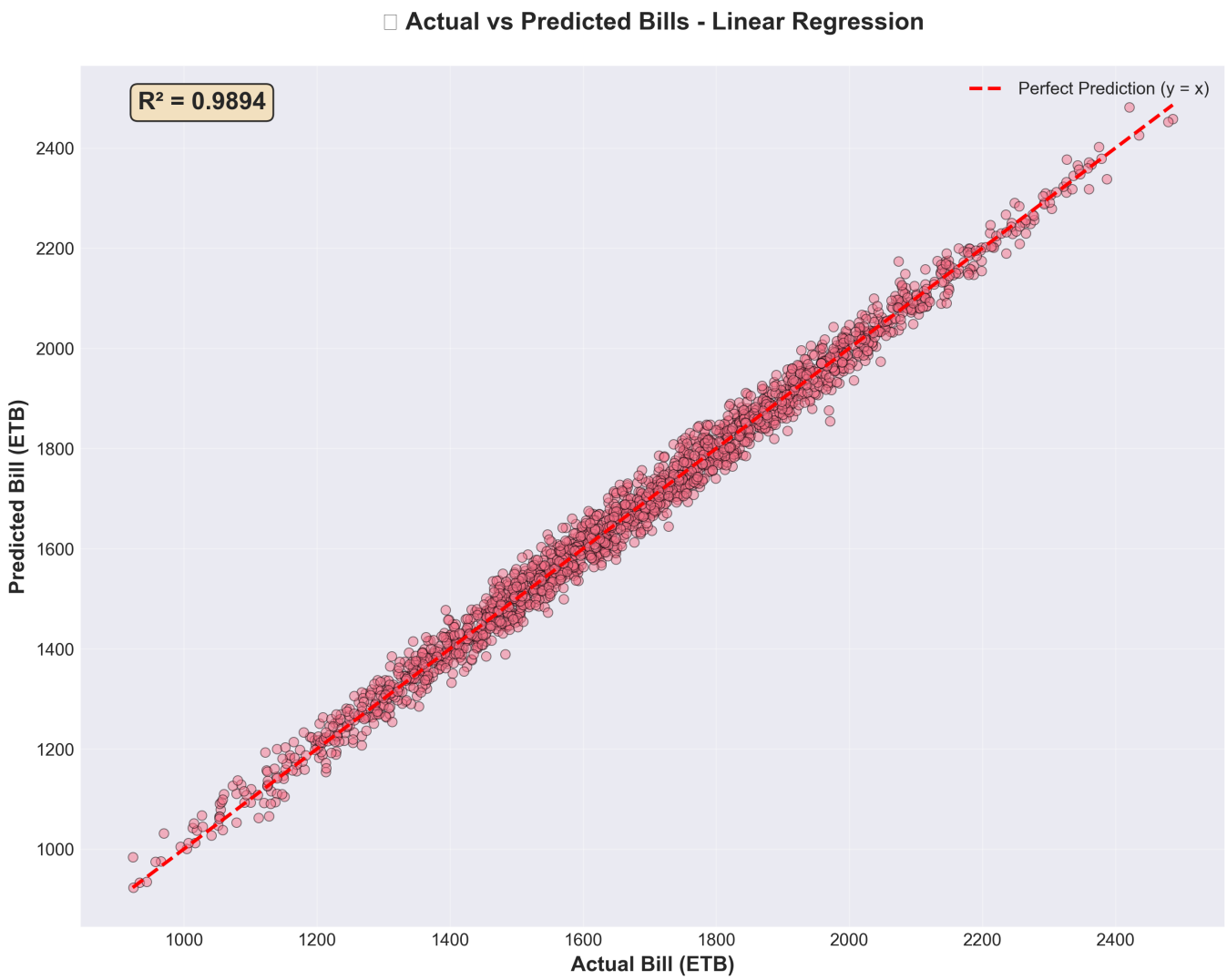
Visualization 1



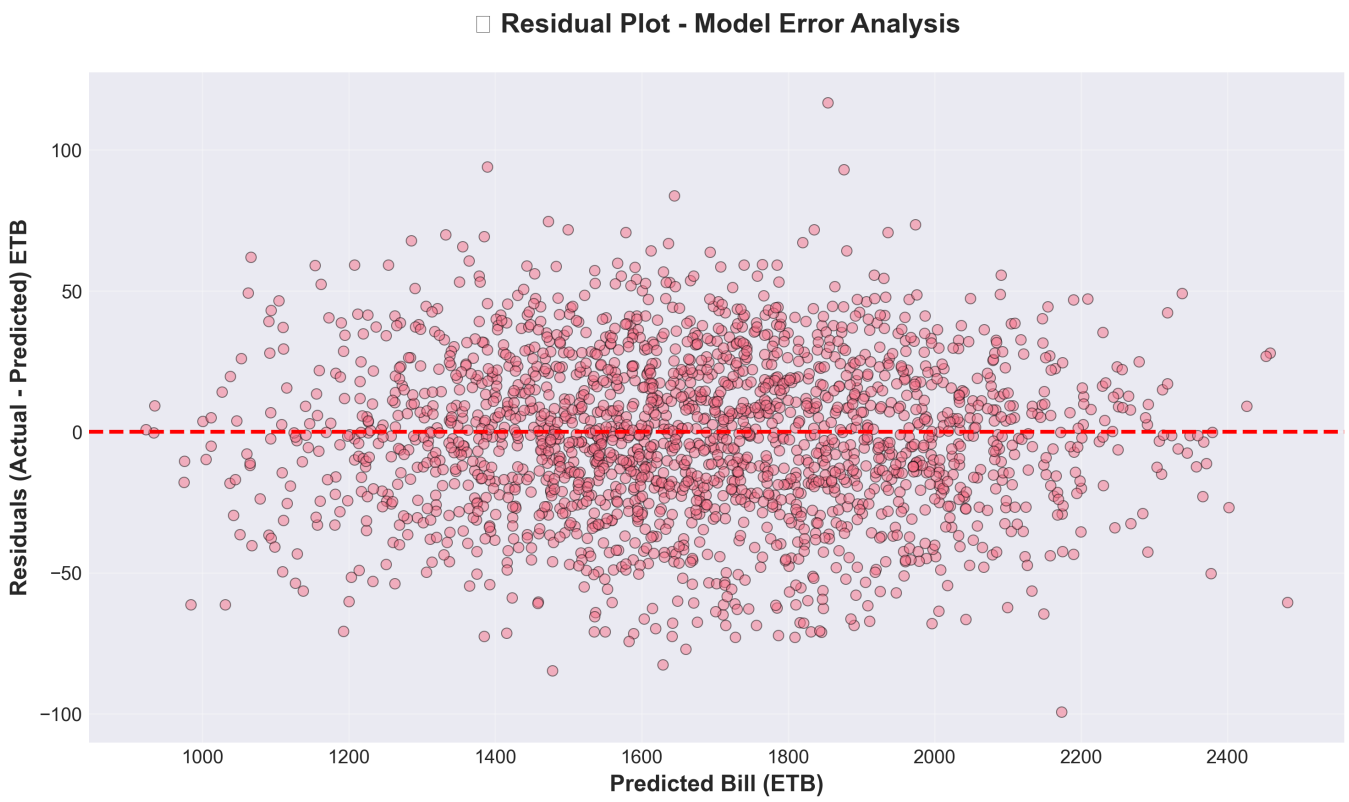
Visualization 2



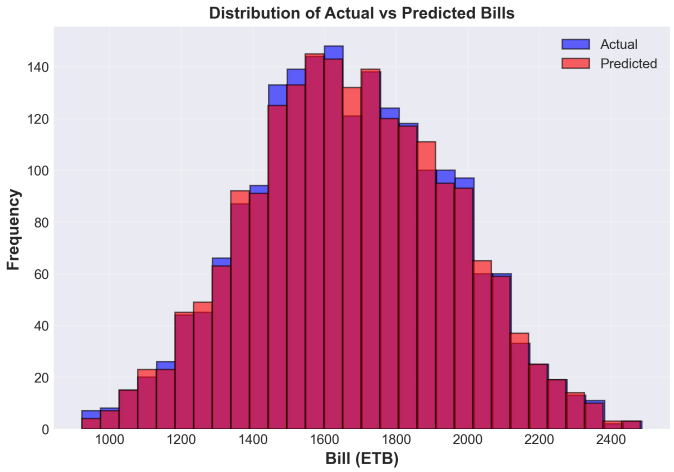
Visualization 3

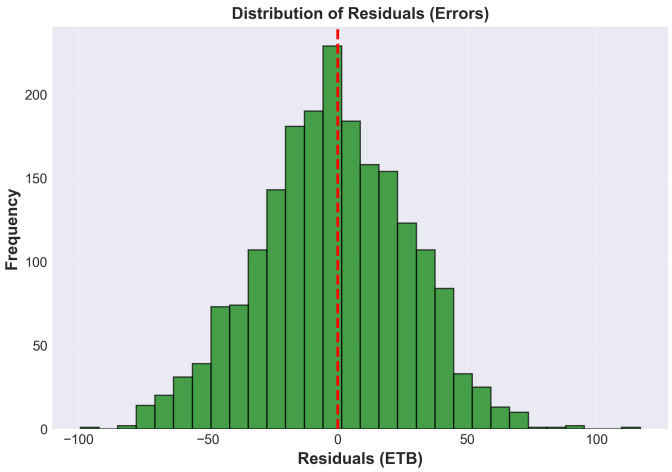


Visualization 4

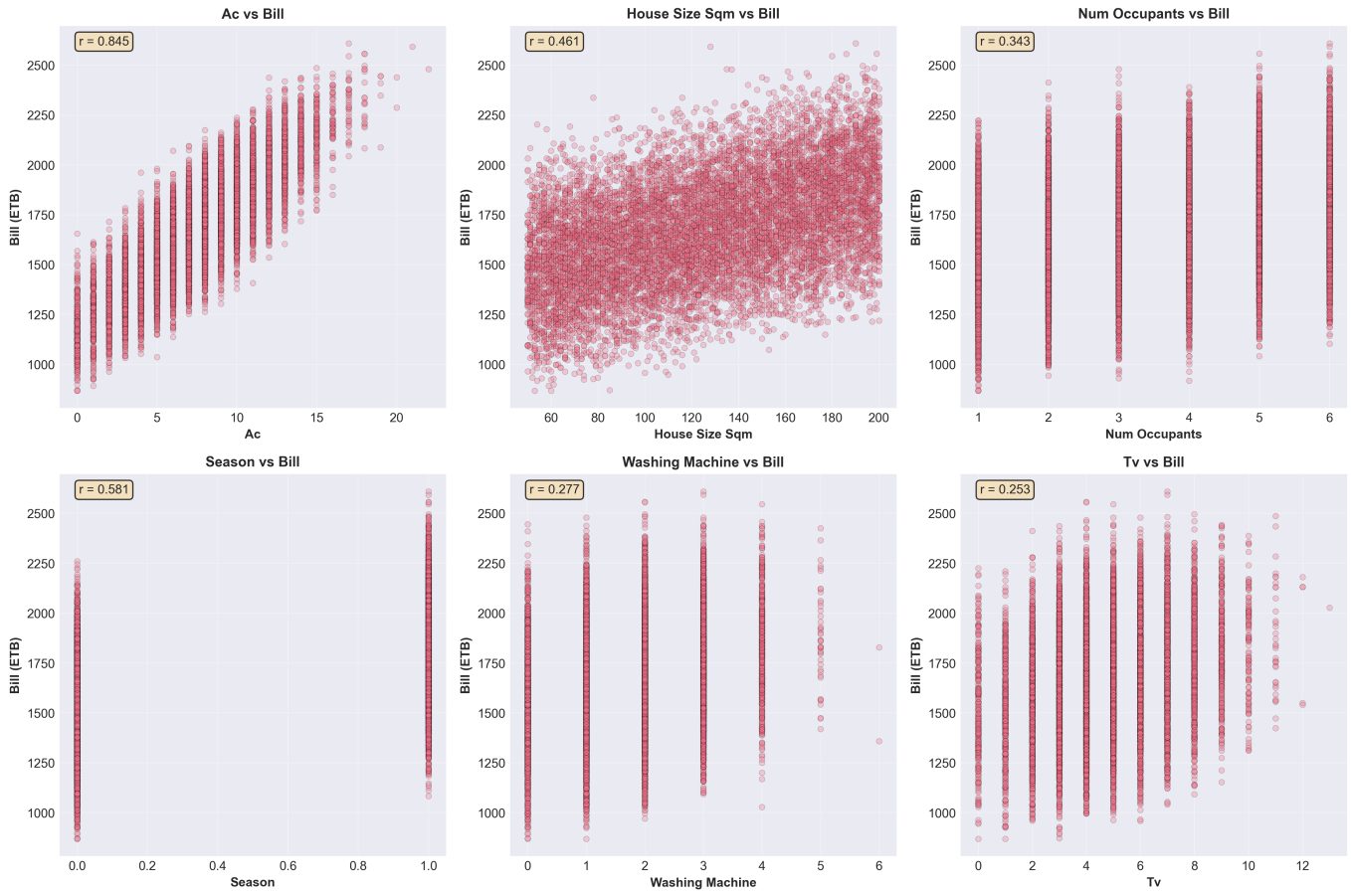


Visualization 5

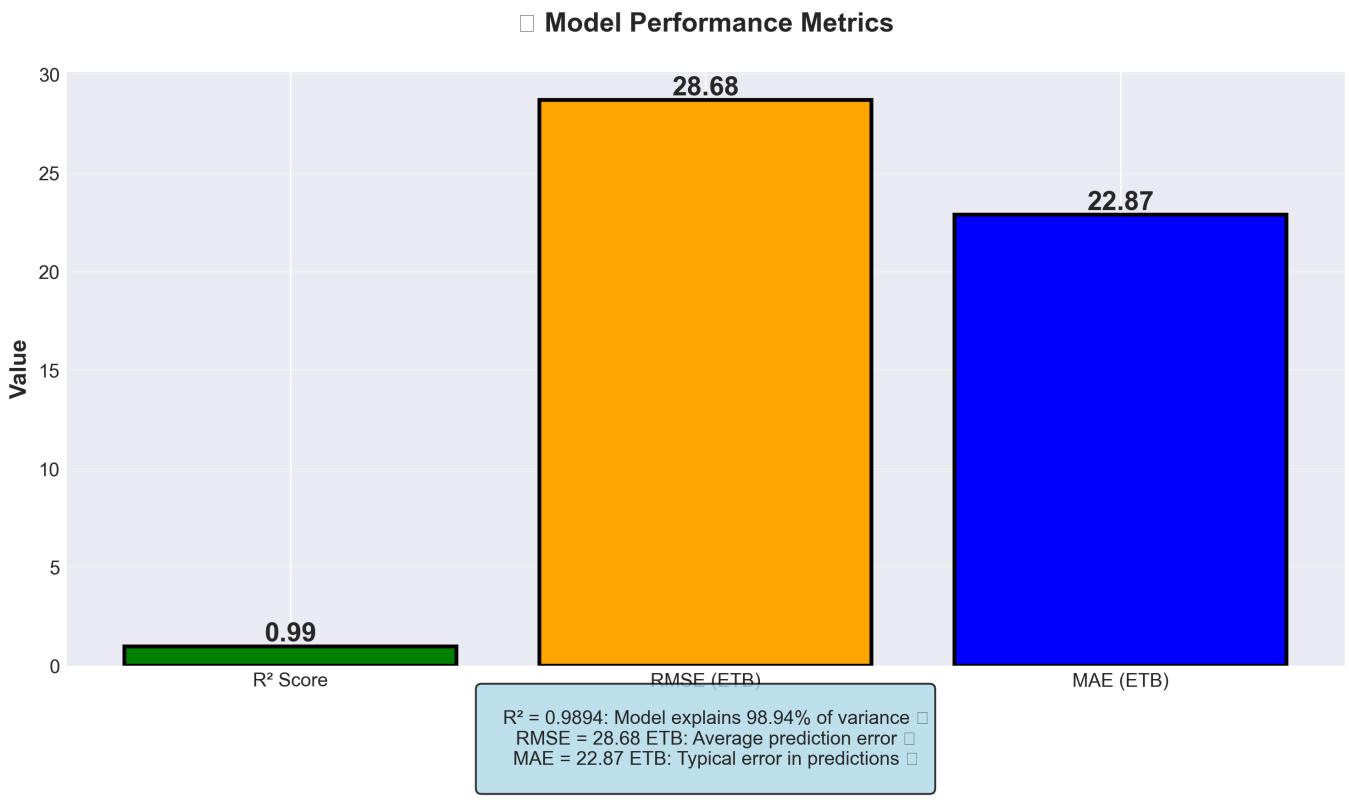




Visualization 6



Visualization 7



6. Conclusion

6.1 Achievements

- Successfully built Linear Regression model for electricity bill prediction
- Achieved exceptional accuracy: R-squared = 0.9894, MAE = 22.87 ETB
- Identified AC as the primary driver of electricity costs
- Created 7 comprehensive visualizations for model interpretation

6.2 K

- AC usage is the most significant factor affecting bills (coefficient: 182.07)
- House size and number of occupants are important secondary factors
- Season affects bills through increased AC and fan usage
- Fridge has minimal impact due to constant, efficient operation

6.3 P

- Households can estimate monthly bills before receiving them
- Identify which appliances to reduce for maximum cost savings
- Budget planning for different seasons (hot vs cool)
- Energy conservation awareness and education

6.4 F

- Add more features: solar panels, insulation quality, building age
- Collect real Ethiopian household data for validation
- Test polynomial regression for potential non-linear relationships
- Build web or mobile interface for easier public access