Aquaculture Machine Learning Analysis

Fish Weight Prediction Model

Linear Regression with PCA Analysis

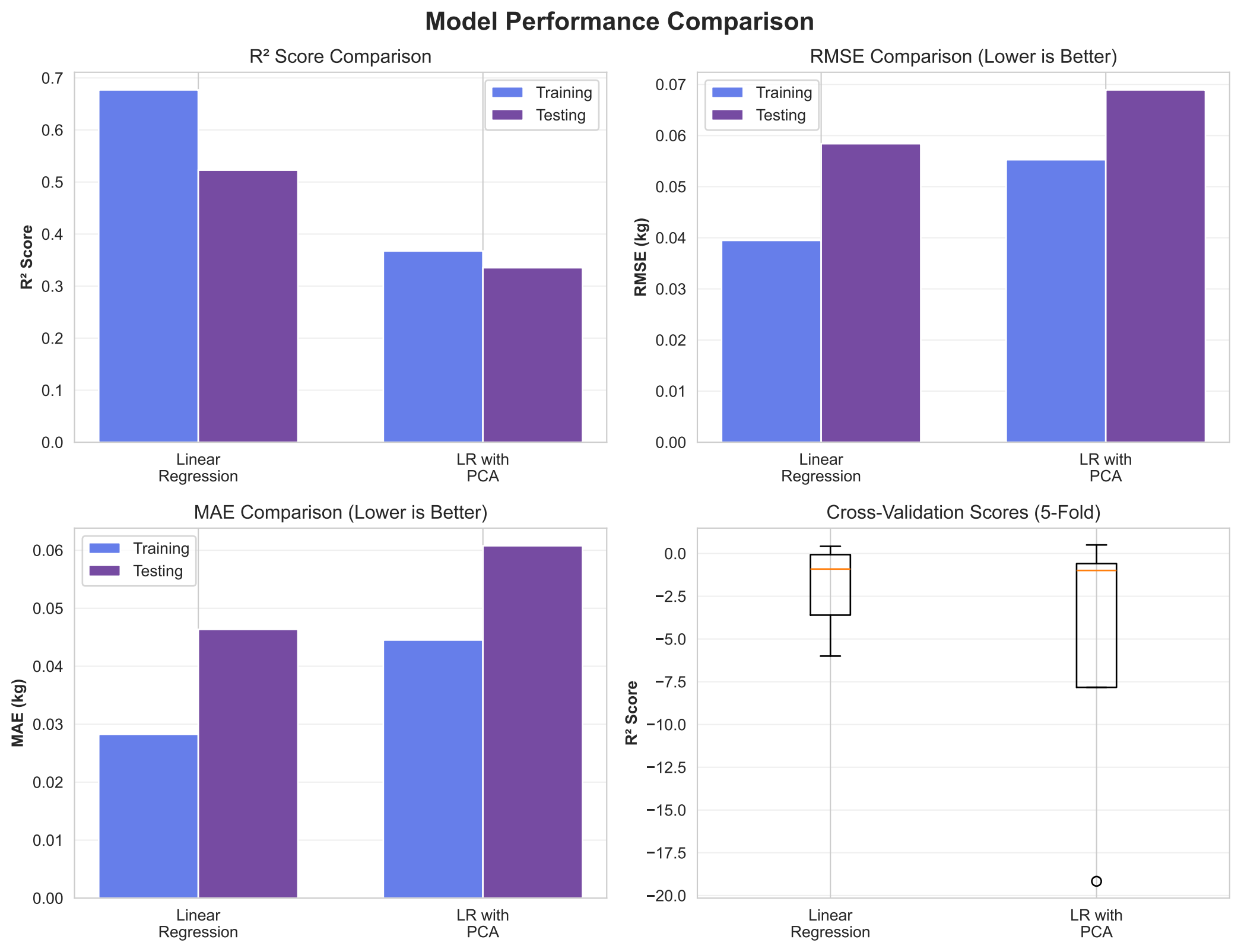
Generated: February 05, 2026 at 09:12 PM

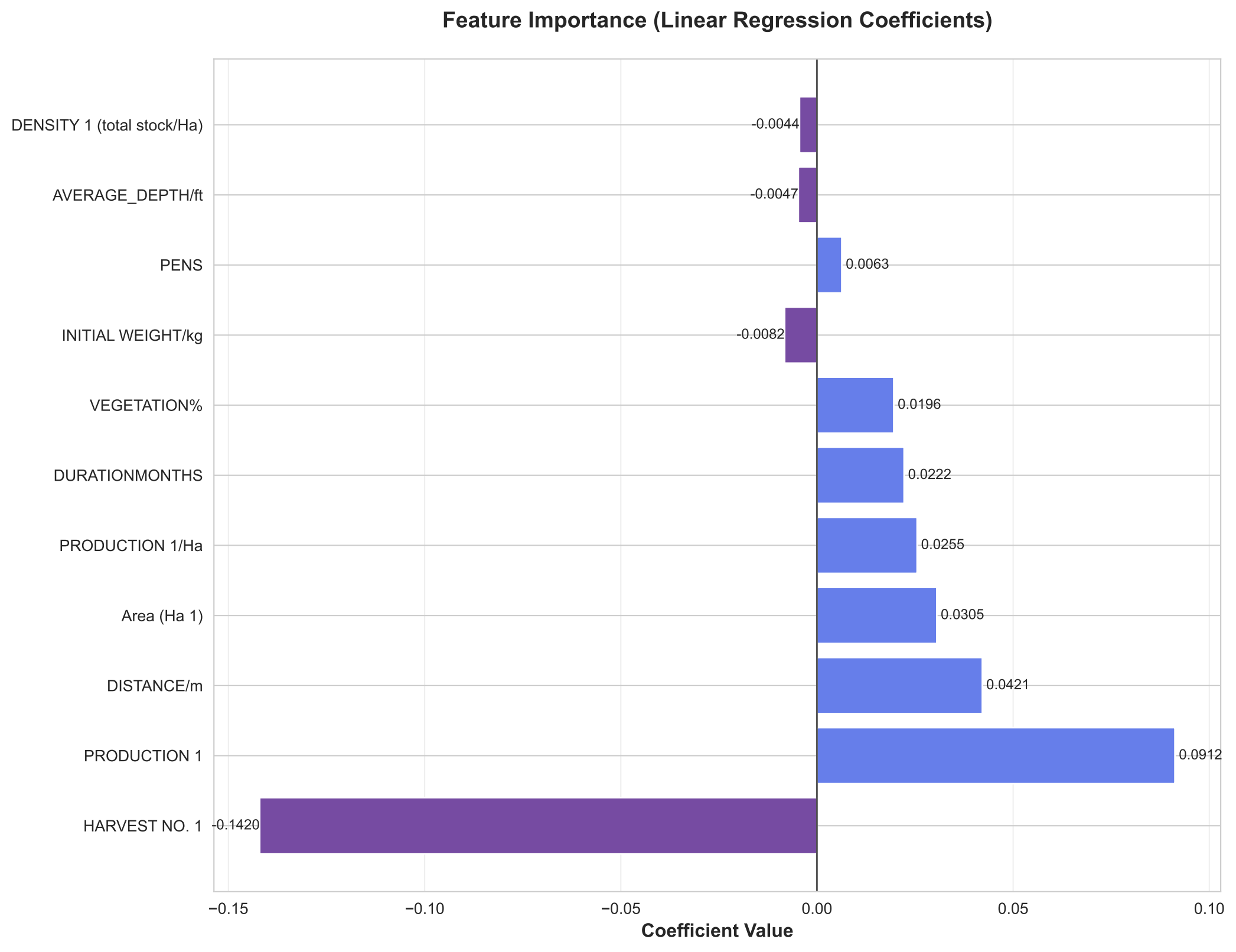
Dataset: 39 samples | 11 features

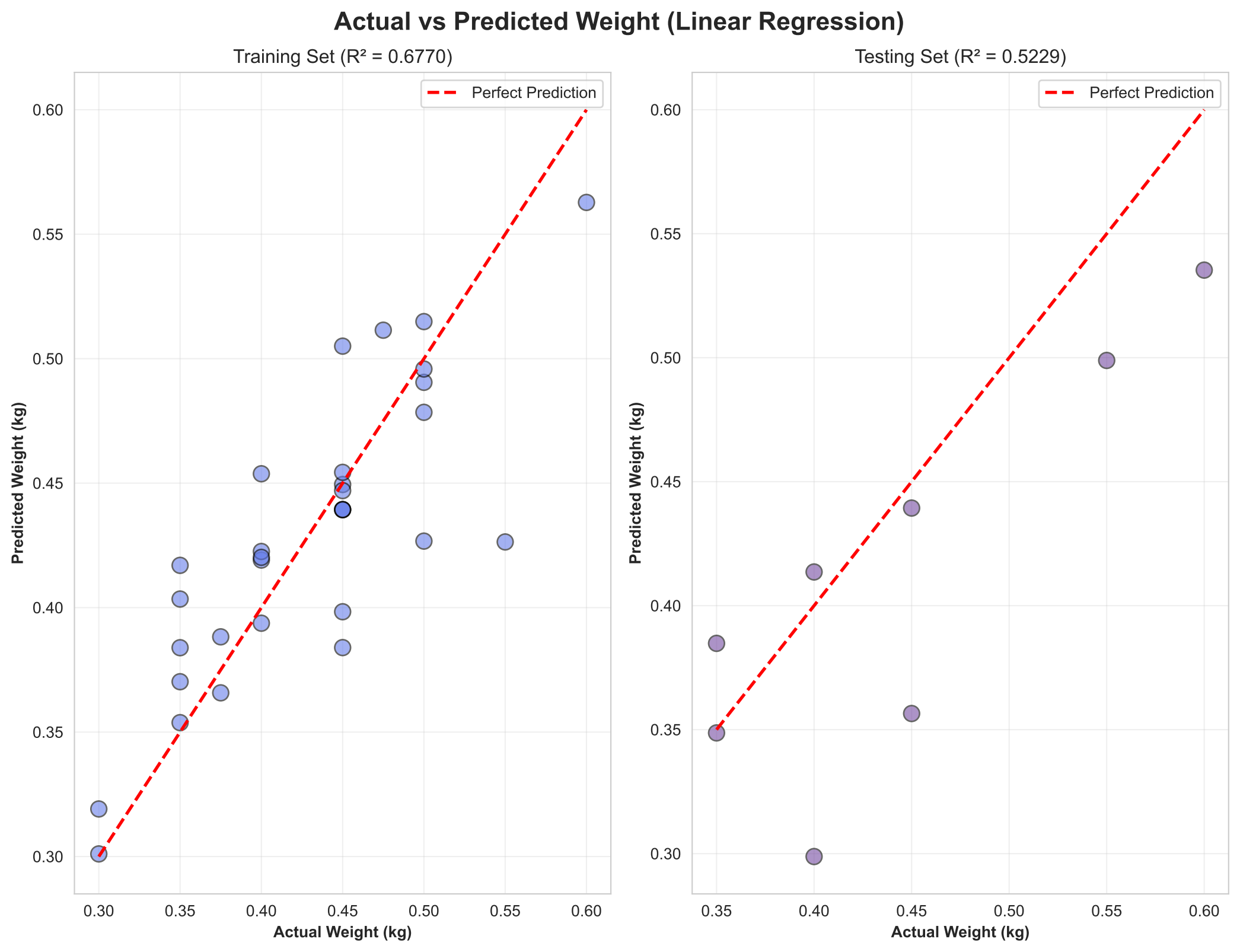
Target Variable: AVG WEIGHT 1 (Fish Final Weight)

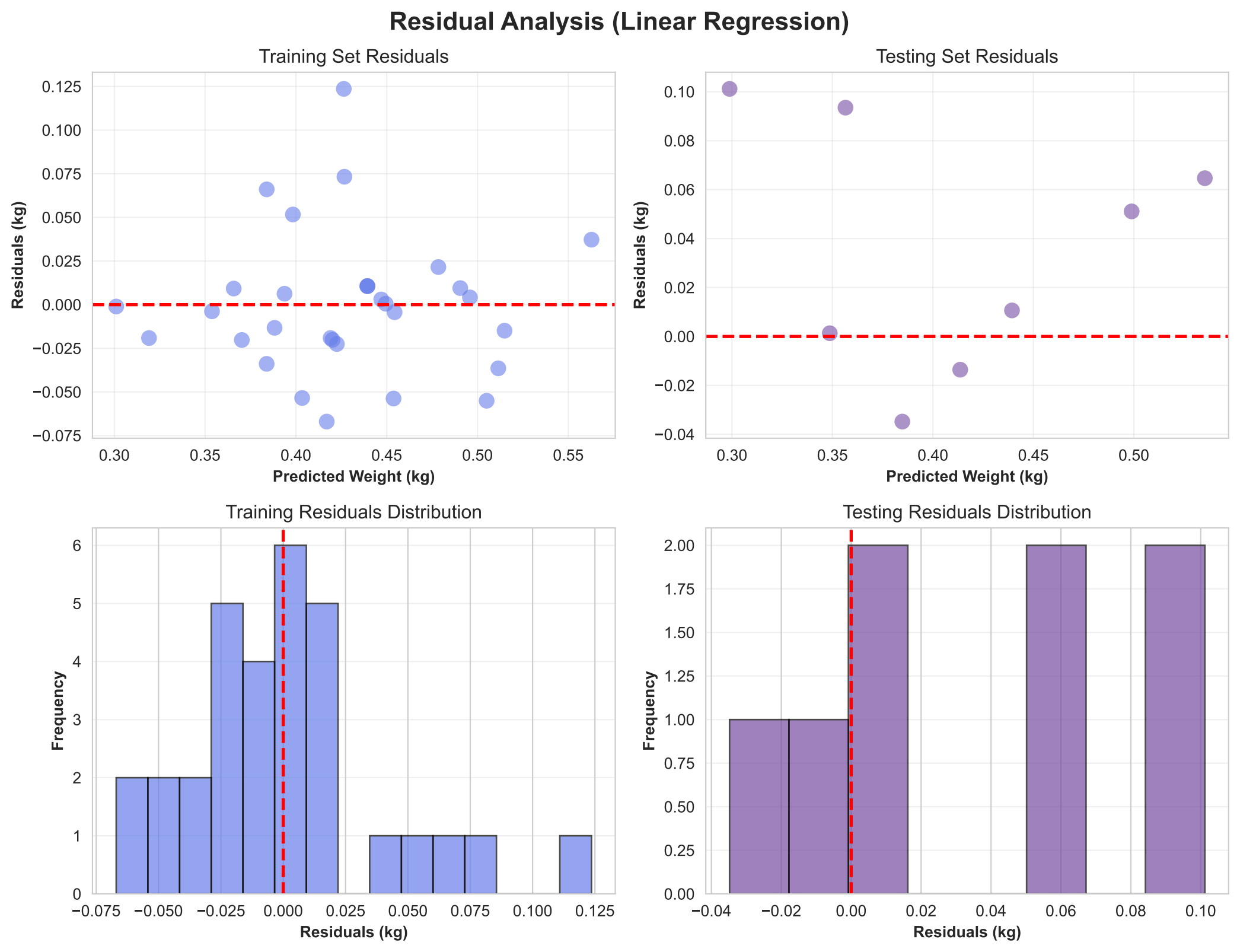
Executive Summary

```
EXECUTIVE SUMMARY
================================================================================

Dataset Overview
  • Total Samples: 39
  • Training Set: 31 samples (79.5%)
  • Testing Set: 8 samples (20.5%)
  • Features: 11
  • Target: AVG WEIGHT 1 (Fish Final Weight)

Target Variable Statistics
  • Range: 0.300 - 0.600 kg
  • Mean: 0.431 kg
  • Standard Deviation: 0.074 kg
  • Median: 0.450 kg

Model 1: Linear Regression (All Features)
  • Training R²: 0.6770 (67.70% variance explained)
  • Testing R²: 0.5229 (52.29% variance explained)
  • Cross-Validation R²: -2.0308 ± 2.4222
  • RMSE: 0.0584 kg
  • MAE: 0.0464 kg

Model 2: Linear Regression with PCA
  • Components: 7 (from 11 features)
  • Variance Retained: 95.44%
  • Testing R²: 0.3354 (33.54% variance explained)
  • RMSE: 0.0689 kg
  • MAE: 0.0608 kg

Best Model: Linear Regression
  • Achieves 52.29% accuracy on test set
  • Average prediction error: ±0.0464 kg
  • Higher accuracy with all features

Top 3 Most Important Features
  1. HARVEST NO. 1: -0.1420
  2. PRODUCTION 1: 0.0912
  3. DISTANCE/m: 0.0421

Key Insight
  The most influential factor is HARVEST NO. 1 with a
  negative correlation,
  suggesting that lower
  values lead to decreased fish weight.
```

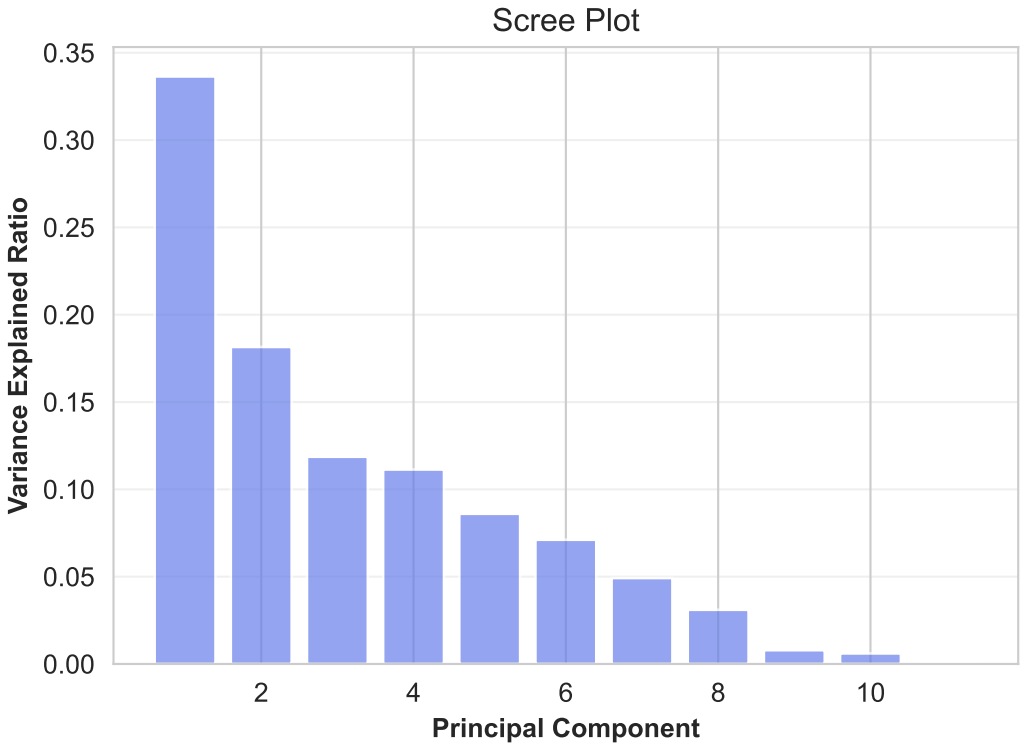

Model Performance Comparison
R² Score Comparison
R² Score
Training
Testing
Linear Regression
LR with PCA
RMSE Comparison (Lower is Better)
RMSE (kg)
MAE Comparison (Lower is Better)
MAE (kg)
Cross-Validation Scores (5-Fold)

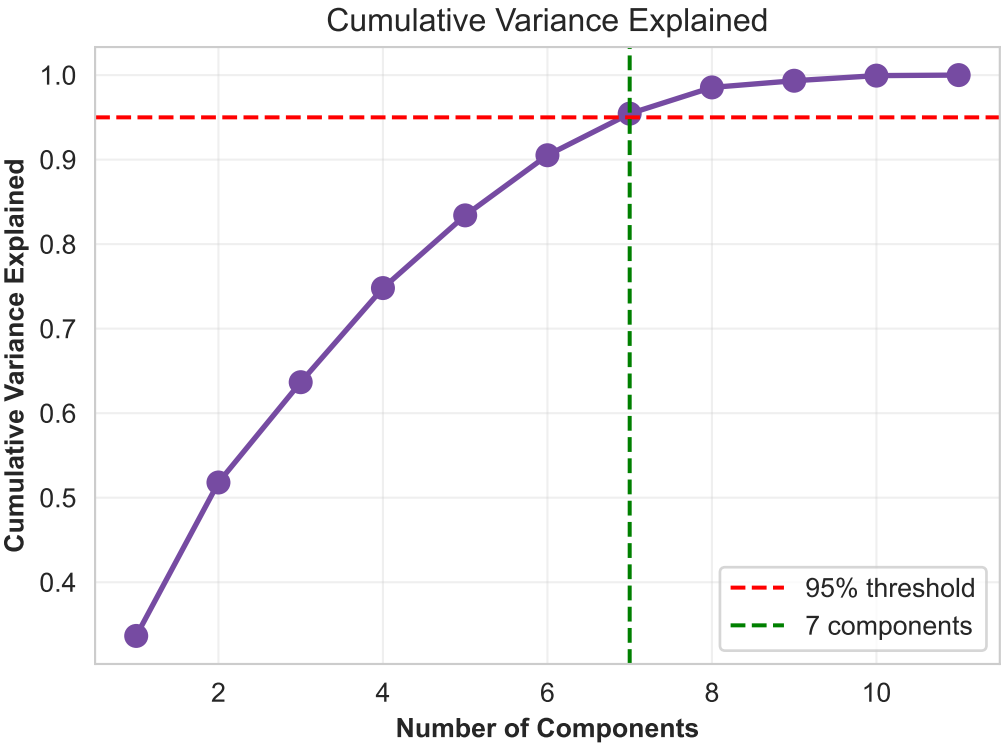

Feature Importance (Linear Regression Coefficients)
DENSITY 1 (total stock/Ha)
AVERAGE_DEPTH/ft
PENS
INITIAL WEIGHT/kg
VEGETATION%
DURATIONMONTHS
PRODUCTION 1/Ha
Area (Ha 1)
DISTANCE/m
PRODUCTION 1
HARVEST NO. 1
-0.0044
-0.0047
0.0063
-0.0082
0.0196
0.0222
0.0255
0.0305
0.0421
0.0912
-0.1420
−0.15
−0.10
−0.05
0.00
0.05
0.10
Coefficient Value

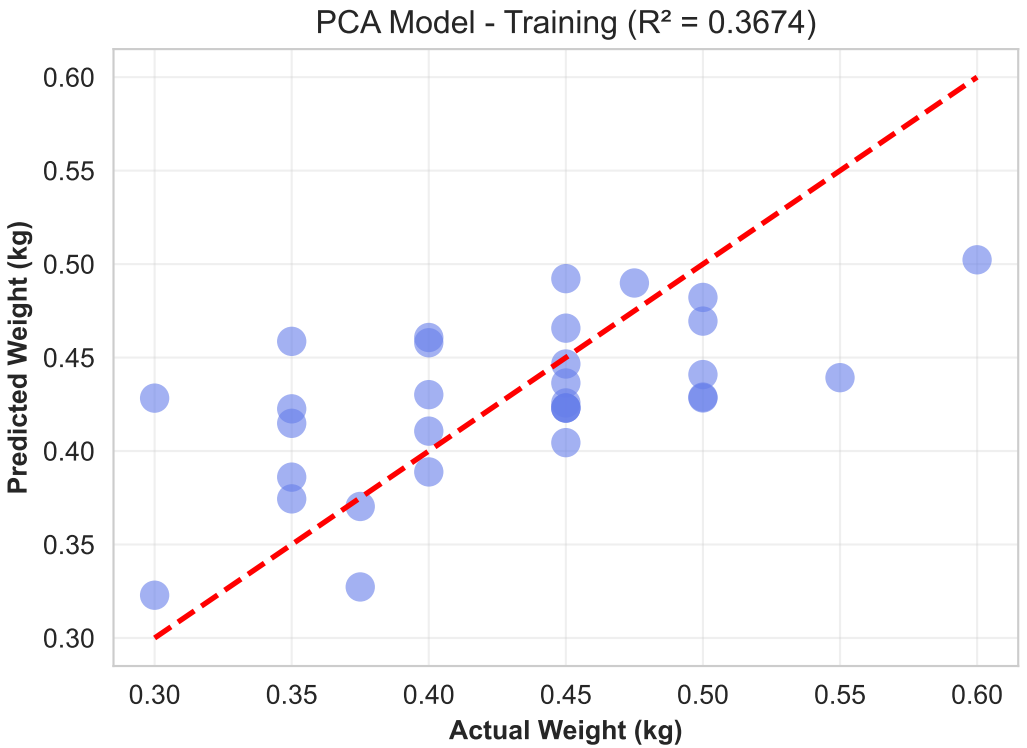

Actual vs Predicted Weight (Linear Regression)
Training Set (R² = 0.6770)
Testing Set (R² = 0.5229)
Perfect Prediction
Predicted Weight (kg)
Actual Weight (kg)
0.30
0.35
0.40
0.45
0.50
0.55
0.60

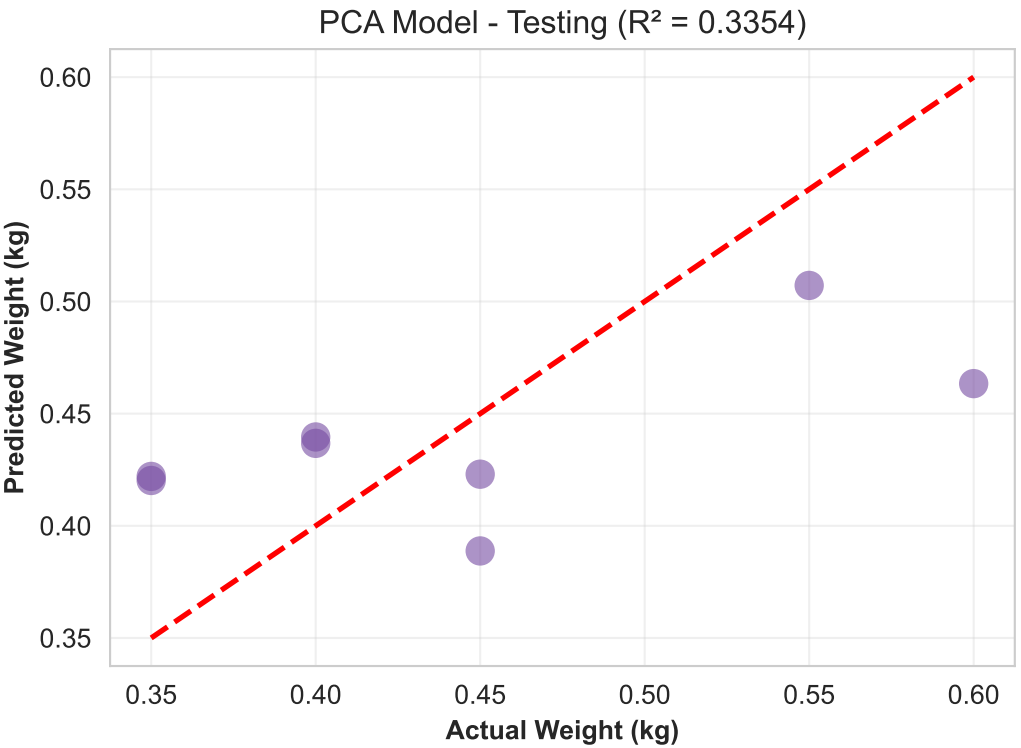

Residual Analysis (Linear Regression)
Training Set Residuals
Residuals (kg)
Predicted Weight (kg)
Testing Set Residuals
Residuals (kg)
Predicted Weight (kg)
Training Residuals Distribution
Frequency
Residuals (kg)
Testing Residuals Distribution
Frequency
Residuals (kg)

Principal Component Analysis









Test Set Predictions

Sample	Actual	LR Pred	LR Error	PCA Pred	PCA Error
Test 1	0.450	0.439	+0.011	0.423	+0.027
Test 2	0.550	0.499	+0.051	0.507	+0.043
Test 3	0.350	0.349	+0.001	0.420	-0.070
Test 4	0.400	0.299	+0.101	0.437	-0.037
Test 5	0.450	0.356	+0.094	0.389	+0.061
Test 6	0.350	0.385	-0.035	0.422	-0.072
Test 7	0.400	0.414	-0.014	0.440	-0.040
Test 8	0.600	0.535	+0.065	0.463	+0.137

Recommendations & Conclusions

RECOMMENDATIONS & CONCLUSIONS
==

Best Model Selection
  ✓ Recommended Model: Linear Regression
  ✓ Test R^2 Score: 0.5229 (52.29% accuracy)
  ✓ Average Error: ±0.0464 kg (46.4 grams)
  ✓ Explains 52.29% of fish weight variation

Model Interpretation
  The model indicates that fish final weight is primarily influenced by:

  1. HARVEST NO. 1 (Coefficient: -0.1420)
     ↓ Higher values → Lighter fish

  2. PRODUCTION 1 (Coefficient: 0.0912)
     ↑ Higher values → Heavier fish

  3. DISTANCE/m (Coefficient: 0.0421)
     ↑ Higher values → Heavier fish

Practical Applications
  • Predict final fish weight based on farming conditions
  • Optimize stocking density and feeding strategies
  • Estimate harvest timing for target weights
  • Identify which factors most impact fish growth

Model Limitations
  • Based on 39 historical samples
  • Predictions most reliable within training data range (0.300 - 0.600 kg)
  • Linear relationship assumed between features and weight
  • External factors (disease, weather) not included

PCA Insights
  • Dimensionality reduced from 11 to 7 features
  • Retained 95.44% of information
  • Simpler model with acceptable accuracy trade-off
  • Consider PCA model for deployment if simplicity is priority

Next Steps
  ✓ Deploy Linear Regression in production
  ✓ Monitor predictions on new data
  ✓ Retrain periodically with updated samples
  ✓ Consider ensemble methods for improved accuracy
  ✓ Collect more data for better generalization

Data Quality Notes
  • 4 samples excluded due to missing values
  • DENSITY column had 0 #REF! errors (cleaned)
  • Consider data collection improvements for future cycles

Report Generated: February 05, 2026 at 09:12 PM
Analysis Tool: Python with scikit-learn
Models: Linear Regression, PCA, StandardScaler