Predictive Modeling Strategy & Hypothesis Testing Framework for Food Processing

Document Type: Strategic Planning Report
Subject Area: Predictive Modeling & Hypothesis Testing for Food Quality
Prepared For: Data Science Teams & Food Processing Organizations
Document Date: December 2025
Version: 1.0

Executive Summary

Building on Week 3's feature engineering foundation, Week 4 focuses on **predictive modeling strategy** and **hypothesis testing.** This document outlines a comprehensive framework for:

- **Problem Definition:** Selecting a critical food processing prediction challenge
- **Hypothesis Development:** Formulating 6 testable hypotheses about quality drivers
- **Model Selection:** Choosing appropriate algorithms (regression, classification, ensemble methods)
- **Evaluation Metrics:** Defining success criteria and validation procedures
- **Implementation Roadmap:** Step-by-step execution plan (Weeks 4–8)

PART 1: PROBLEM DEFINITION

1.1 Selected Prediction Problem: Shelf-Life Prediction (Regression)

Problem Statement

Objective: Develop a predictive model that accurately estimates the remaining shelf-life (in days) of food products based on **process parameters, environmental conditions, raw material characteristics, and operational variables**.

Business Context:

- Food processing organizations currently use **generic shelf-life estimates** (e.g., "all yogurt: 21 days at 4°C")
- Reality: Actual shelf-life varies due to **process variations, supplier differences, storage conditions, and equipment performance**
- **Impact:** Premature product rejection (waste, lost revenue) OR overestimated shelf-life (spoilage, food safety risk)

Why Shelf-Life Prediction?

Business Value (ROI):

1. **Waste Reduction:** Accurate shelf-life → fewer premature rejections → cost savings 5–15%
2. **Food Safety:** Prevent spoilage incidents → reduce recall risk, regulatory fines
3. **Supply Chain Optimization:** Predict shelf-life at production → optimize distribution, reduce storage cost
4. **Customer Satisfaction:** Fresher products delivered → reduced complaints, improved brand reputation

Target Variable Definition

Shelf-Life (days): Time from production completion until the product reaches unacceptable quality (microbial load exceeds limit OR sensory attributes degrade to threshold OR chemical indicators exceed limit).

Range: 5–45 days (typical for dairy, bakery, beverage products)

1.2 Problem Scope & Constraints

Scope Boundaries

In Scope:

- Single production facility OR aggregated across similar facilities
- Specific product category (e.g., yogurt, cheese, jam, juice)
- Historical data: 12–36 months (sufficient seasonal variation)
- Features: Temperature, humidity, pH, microbial counts, supplier ID, production line, raw material age, batch characteristics

Data Requirements

Requirement	Specification
Sample Size	≥500 batches (preferably 1,000+)
Time Period	≥12 months (ideally 24–36)
Target Variable	Actual shelf-life (measured, not estimated)
Feature Completeness	≥90% of rows have process data; <5% missing critical features
Data Quality	No spurious outliers; validated sensor readings

PART 2: HYPOTHESIS FORMULATION

2.1 Hypothesis Development Framework

Hypothesis Definition: A testable prediction about the relationship between predictor variables and shelf-life outcome.

Format: "IF [condition on predictor], THEN [expected direction/magnitude of effect on shelf-life], BECAUSE [mechanistic reasoning]"

2.2 Six Core Hypotheses

Hypothesis 1: Temperature Control Dominates Shelf-Life

Statement: Batches processed and stored at lower, more stable temperatures have significantly longer shelf-life.

Reasoning:

- **Enzymatic Degradation:** Shelf-life degradation follows Arrhenius kinetics → exponential acceleration with temperature rise
- **Microbial Growth:** Optimal growth temperature: 20–35°C. Below 10°C: growth minimal. Above 45°C: most pathogens inhibited.
- **Expected Effect:** +2–3 days shelf-life per 1°C reduction in average storage temperature (within 5–30°C range)
- **Predictive Power:** Temperature likely top 3 most important features in model

Mechanistic Support:

- Q_{10} Rule: Reaction rate doubles per 10°C increase
- For shelf-life: Every 5°C ↑ → shelf-life ÷ 2 to ÷ 3 (approximately)

Measurement: Regression coefficient for temperature features should be significant ($p<0.05$) and negative

---

Hypothesis 2: pH & Acidity Provide Antimicrobial Barriers

Statement: Products with lower pH (more acidic; pH < 4.0) exhibit longer shelf-life because acidity inhibits pathogenic and spoilage microorganisms.

Reasoning:

- **Antimicrobial Effect:** Low pH (high H^+ concentration) denatures microbial cell membranes
- **Regulatory Relevance:** FDA classifies foods pH < 4.6 as "acidic foods" → reduced botulism risk
- **Expected Effect:**
  - pH 3.0–4.0: Highly resistant to spoilage → shelf-life ≥25 days
  - pH 4.0–5.5: Moderate risk → shelf-life 15–20 days
  - pH > 5.5: High risk → shelf-life <15 days

Mechanistic Support:

- Clostridium botulinum inactivated below pH 4.6
- Salmonella inhibited below pH 3.8
- Most spoilage yeasts/molds inhibited below pH 4.0

Measurement: Regression coefficient for pH should be significant and negative

---

Hypothesis 3: Microbial Load at Production Predicts Shelf-Life Trajectory

Statement: Batches with lower initial microbial contamination achieve longer shelf-life, even under suboptimal storage.

Reasoning:

- **Growth Dynamics:** If starting count is 100 CFU/mL vs. 10,000 CFU/mL, first batch reaches spoilage threshold hours/days later
- **Exponential Growth:** $N(t) = N_0 \times e^{(\mu t)}$, where μ = growth rate
- **Spoilage Time:** $t = \ln(N_{limit} / N_0) / \mu \rightarrow$ lower $N_0 \rightarrow$ longer t
- **Expected Effect:** 1 order of magnitude (10×) reduction in initial CFU → ~2–3 days longer shelf-life

Mechanistic Support:

- Microbial growth follows exponential kinetics at constant conditions
- Time-to-threshold is logarithmically dependent on initial count

Measurement: Log-transformed microbial feature more predictive than raw count

---

Hypothesis 4: Humidity Drives Mold & Oxidation Risks

Statement: Products stored in high-humidity environments (>70% RH) experience accelerated quality degradation, reducing shelf-life.

Reasoning:

- **Mold Germination:** Requires water activity (a_w) > 0.75–0.90
- **Humidity-Water Activity Link:** High RH → higher product equilibrium moisture → higher a_w
- **Oxidation:** Humidity increases browning reactions, nutrient loss, rancidity
- **Expected Effect:** Every 10% RH increase above 60% → -1 to -2 days shelf-life

Mechanistic Support:

- Aspergillus, Penicillium mold germination threshold: RH ≈ 80–90%
- Sorption isotherms: RH determines equilibrium moisture content

Measurement: Regression coefficient negative and significant for humidity

---

Hypothesis 5: Supplier Quality Score Predicts Batch Performance

Statement: Raw materials from suppliers with historically higher quality scores produce finished products with longer shelf-life.

Reasoning:

- **Incoming Quality Impact:** Contamination in raw materials can't be fully eliminated by processing
- **Supplier Variability:** Some suppliers maintain tighter quality standards → lower incoming microbial load
- **Expected Effect:** Supplier quality score correlates 0.3–0.5 with finished product shelf-life

Mechanistic Support:

- Raw material microbial load is starting point for finished product
- Supplier certifications (ISO 22000, FSSC 22000) correlate with product consistency

Measurement: Regression coefficient positive; shows supplier quality drives shelf-life

---

Hypothesis 6: Equipment Performance & Maintenance State Affects Processing Quality

Statement: Batches produced on well-maintained equipment with stable operating parameters have longer shelf-life.

Reasoning:

- **Equipment Drift:** Over time, sensors drift, gaskets leak, thermal efficiency degrades
- **Process Deviation:** Drifting temperature → inadequate thermal treatment → insufficient pathogen reduction
- **Stability Index:** Equipment with low temperature variance → better process control → higher final product quality
- **Expected Effect:** Equipment stability index correlates 0.4–0.6 with shelf-life

Mechanistic Support:

- Thermal processing depends on precise temperature-time combinations
- Equipment degradation → under-processing → inadequate microbial reduction

Measurement: Regression coefficient positive (higher stability → longer shelf-life)

PART 3: PREDICTIVE MODEL SELECTION

3.1 Why Regression (Not Classification)?

Classification Approach (e.g., "Good/Bad"):

- ✗ Loses information (20-day vs. 25-day both classified as "Long")
- ✗ Threshold selection is arbitrary
- ✗ Business needs continuous prediction

Regression Approach (e.g., predict exact shelf-life: 18.5 days):

- ✓ Captures continuous variation
- ✓ Enables fine-grained decisions
- ✓ Supports probabilistic forecasting

Decision: Shelf-life prediction will use **regression models** as primary approach.

3.2 Candidate Models (Conceptual Framework)

Model 1: Multiple Linear Regression (MLR)

Concept: Shelf-Life = $\beta_0 + \beta_1 \times \text{Temp} + \beta_2 \times \text{Humidity} + \beta_3 \times \text{pH} + \ldots + \varepsilon$

Advantages:

- ✓ Fully interpretable; each coefficient has direct meaning
- ✓ Assumptions transparent (residual analysis reveals violations)
- ✓ Fast to train; minimal hyperparameter tuning
- ✓ Ideal for hypothesis testing (p-values, confidence intervals)

Disadvantages:

- ✗ Assumes linear relationships (may miss non-linear patterns)
- ✗ Sensitive to outliers
- ✗ Multicollinearity can inflate coefficients

Expected Performance:

- $R^2 \approx 0.65–0.75$
- RMSE ≈ 2.5–3.5 days

Model 2: Decision Tree Regression

Concept: Recursively split feature space to minimize prediction variance.

Advantages:

- ✓ Handles non-linear relationships naturally
- ✓ Captures interaction effects implicitly
- ✓ Interpretable (decision rules understandable to stakeholders)
- ✓ No feature scaling required

Disadvantages:

- ✗ Prone to overfitting
- ✗ Small data changes can drastically change tree structure
- ✗ Greedy algorithm (local optimization)

Expected Performance:

- $R^2 \approx 0.70–0.80$
- RMSE ≈ 2.0–3.0 days

Model 3: Random Forest Regression

Concept: Train 100–500 decision trees on random subsets; average predictions.

Advantages:

- ✓ Reduces overfitting risk
- ✓ Handles non-linearities and interactions
- ✓ Robust to outliers
- ✓ Provides feature importance ranking
- ✓ Superior predictive accuracy
- ✓ Parallelizable (fast training)

Disadvantages:

- ✗ Less interpretable ("black box")
- ✗ Requires more hyperparameter tuning
- ✗ Slower inference than linear models

Expected Performance:

- $R^2 \approx 0.75–0.85$ (best overall)
- RMSE ≈ 1.5–2.5 days

Model 4: Gradient Boosting Machine (GBM) / XGBoost

Concept: Sequentially train trees, each correcting errors of previous trees.

Advantages:

- ✓ Often highest predictive accuracy
- ✓ Handles non-linearities and interactions
- ✓ Robust to outliers and imbalanced data
- ✓ Feature importance well-calibrated

Disadvantages:

- ✗ Complex hyperparameter tuning required
- ✗ Risk of overfitting
- ✗ Slower training than random forest
- ✗ Difficult to interpret

Expected Performance:

- $R^2 \approx 0.80–0.88$ (highest accuracy)
- RMSE ≈ 1.2–2.0 days

Model 5: Support Vector Regression (SVR)

Concept: Find hyperplane that minimizes prediction error within ε-margin.

Advantages:

- ✓ Excellent for small-to-medium datasets

- ✓ Handles high-dimensional feature spaces
- ✓ Kernel tricks enable complex non-linear relationships

Disadvantages:

- ✗ Requires feature scaling
- ✗ Hyperparameter tuning complex
- ✗ Not interpretable
- ✗ Slower prediction than tree-based methods

Expected Performance:

- $R^2 \approx 0.70–0.82$
- RMSE ≈ 1.8–2.8 days

3.3 Recommended Model Strategy

Primary Model: Random Forest Regression

Justification:

1. **Balance:** Good accuracy + reasonable interpretability
2. **Robustness:** Handles outliers, non-linearities, interactions naturally
3. **Generalization:** Ensemble approach reduces overfitting
4. **Hyperparameter Tuning:** Well-understood
5. **Production-Ready:** Fast inference; scales well

Configuration:

```
RandomForestRegressor(
n_estimators=200,
max_depth=10,
min_samples_leaf=15,
min_samples_split=30,
max_features='sqrt',
random_state=42
)
```

Secondary Model: Linear Regression (Interpretability)

Justification:

- Hypothesis testing requires interpretable coefficients
- Baseline for comparison
- Regulatory/stakeholder reporting

Validation Approach: Ensemble Voting

Concept: Train 3–4 models; make predictions as weighted average.

Example Ensemble:

- 40% Random Forest weight
- 30% Gradient Boosting weight

- 20% Linear Regression weight
- 10% SVR weight

PART 4: EVALUATION METRICS & VALIDATION FRAMEWORK

4.1 Regression Evaluation Metrics

Metric 1: Root Mean Squared Error (RMSE)

Formula: $RMSE = \sqrt{((1/n) \times \Sigma(y\_pred - y\_actual)^2)}$

Interpretation:

- Units: Days (same as target)
- Example: RMSE = 2.5 days → average prediction error is ±2.5 days
- Lower is better
- **Target:** RMSE < 2.5 days (acceptable), < 2.0 days (excellent)

Metric 2: Mean Absolute Error (MAE)

Formula: $MAE = (1/n) \times \Sigma|y\_pred - y\_actual|$

Interpretation:

- Units: Days
- Example: MAE = 1.8 days → average prediction error magnitude is 1.8 days
- Lower is better
- **Target:** MAE < 2.0 days

Metric 3: R^2 (Coefficient of Determination)

Formula: $R^2 = 1 - (SS\_res / SS\_tot)$

Interpretation:

- Range: 0–1 (1 = perfect predictions)
- Example: $R^2 = 0.78$ → model explains 78% of variance in shelf-life
- Higher is better
- **Target:** $R^2 > 0.75$ (good), > 0.85 (excellent)

Metric 4: Mean Absolute Percentage Error (MAPE)

Formula: $MAPE = (1/n) \times \Sigma(|y\_actual - y\_pred| / |y\_actual|) \times 100\%$

Interpretation:

- Units: Percentage
- Example: MAPE = 12% → average percentage error is 12%
- Lower is better

- **Target:** MAPE < 15%

Metric 5: Median Absolute Error (MedAE)

Formula: MedAE = median(|y_actual - y_pred|)

Interpretation:

- Units: Days
- Robust to outliers (uses median, not mean)
- **Target:** MedAE < 2.0 days

Metric 6: Prediction Interval Coverage (95% CI)

Concept: For each prediction, compute 95% confidence interval; check if actual value falls within.

Interpretation:

- Example: Coverage = 0.94 → 94% of actual values fall within predicted 95% CI
- **Target:** Coverage ≈ 0.95 ± 0.03
- Important for production: "Batch shelf-life: 18 days (95% CI: 16–20 days)"

4.2 Validation Framework

Strategy 1: Train-Test Split (Simple Baseline)

Procedure:

1. Randomly split data: 80% train, 20% test
2. Train model on training set
3. Evaluate metrics on test set

When to Use: Quick prototyping, small datasets

Strategy 2: K-Fold Cross-Validation (Recommended)

Procedure:

1. Divide data into k folds (typically k=5)
2. For each fold: Train on k-1 folds; Evaluate on held-out fold
3. Report mean ± std of metrics across folds

Interpretation:

- Mean R^2 = 0.776 ± 0.023 → typical model performance ± variability
- Std R^2 = 0.023 → stable across folds (low variance)
- **When to Use:** Standard validation; recommended for food processing

---

Strategy 3: Time-Series Cross-Validation (For Temporal Data)

Procedure:

1. Exploit temporal ordering of batches
2. For each split: Train on past data; Evaluate on future data

Example:
Split 1: Train on batches 1–100; Test on batches 101–120
Split 2: Train on batches 1–120; Test on batches 121–140
Split 3: Train on batches 1–140; Test on batches 141–160

Advantage: Realistic evaluation (model predicts future); prevents data leakage

When to Use: Production models where temporal order matters (recommended for food processing with seasonal patterns)

4.3 Hypothesis Testing Within Model

Test 1: Feature Significance (Regression Coefficients)

For Linear Regression:

- Extract coefficient for each feature
- Compute p-value (via t-test)
- Features with $p < 0.05$ are statistically significant

Expected Output for Hypotheses:
Feature Coefficient p-value Hypothesis
Temperature (°C) -0.852 <0.001 ✓ H1 Supported
Humidity (%) -0.034 <0.001 ✓ H4 Supported
pH -0.512 <0.001 ✓ H2 Supported
Log Microbial Count -1.234 <0.001 ✓ H3 Supported
Supplier Quality Score 0.512 <0.001 ✓ H5 Supported
Equipment Stability Index 0.456 <0.001 ✓ H6 Supported

Test 2: Feature Importance (Tree-Based Models)

For Random Forest / Decision Tree:

- Compute feature importance (Gini-based, permutation-based)
- Rank features by importance
- Top features should correspond to hypotheses

Expected Output:
Rank Feature Importance Hypothesis
1 Temperature Squared 0.224 ✓ H1
2 Temperature 0.189 ✓ H1
3 Microbial Count (Log) 0.167 ✓ H3
4 Equipment Stability 0.112 ✓ H6
5 pH Distance from Optimal 0.098 ✓ H2

6 Humidity 0.087 ✓ H4
7 Supplier Quality Score 0.065 ✓ H5

Test 3: Residual Analysis

Residuals: Actual shelf-life minus predicted shelf-life for each batch

Tests:

1. **Normality:** Residuals should be approximately normal
    - Shapiro-Wilk test: $p > 0.05$ ✓ Residuals normal
2. **Heteroscedasticity:** Residuals should have constant variance
    - Breusch-Pagan test: $p > 0.05$ ✓ Constant variance
3. **Autocorrelation:** Residuals should be independent
    - Durbin-Watson test: $DW \approx 2$ ✓ No autocorrelation

---

PART 5: IMPLEMENTATION ROADMAP

5.1 Phase-Gated Execution (Weeks 4–8; ~30–35 Hours)

Phase 1: Problem & Hypothesis Validation (Week 4; 5–6 hours)

Objective: Confirm problem framing; validate hypotheses via exploratory analysis.

Day	Task	Duration	Deliverable
1	Confirm problem definition; stakeholder alignment	1 hr	Problem definition signed-off
2–3	Conduct EDA per hypothesis; plot correlations, trends	2.5 hrs	6 hypothesis validation plots
4	Statistical significance testing (t-tests, ANOVA)	1 hr	Statistical test results
5	Hypothesis support scorecard; document findings	0.5 hrs	Hypothesis validation report

Expected Outcome: 4–6 hypotheses supported ($p<0.05$); 0–2 may require refinement.

Phase 2: Data Preparation & Feature Selection (Week 5; 5–6 hours)

Objective: Clean data; select features for modeling.

Day	Task	Duration	Deliverable
1–2	Load raw data; check quality (missing values, outliers)	2 hrs	Data quality report; cleaning log
3	Apply transformations (scaling, log, Box-Cox)	1.5 hrs	Transformed feature set
4	Feature selection: correlation, VIF, p-values	1 hr	Selected feature subset (~15–20 features)
5	Create training/validation/test splits	0.5 hrs	Data splits (80% train, 10% val, 10% test)

Expected Outcome: Ready-to-model dataset with <5% missing; feature multicollinearity (VIF) < 5.

Phase 3: Baseline & Linear Model (Week 5–6; 5–6 hours)

Objective: Train interpretable baseline; validate hypotheses.

Day	Task	Duration	Deliverable
1–2	Train Linear Regression; compute coefficients, p-values	2 hrs	MLR model; coefficient table
3	Generate regression plots; residual analysis	1.5 hrs	Residual plots; diagnostic plots
4	Cross-validation (5-fold); report CV metrics	1 hr	CV results (R^2, RMSE, MAE)
5	Hypothesis validation via coefficients; document	0.5 hrs	Hypothesis support scorecard

Expected Outcome:

- Baseline $R^2 \approx 0.65–0.75$
- RMSE ≈ 2.5–3.5 days
- ≥5 hypotheses show significant coefficients

Phase 4: Advanced Model Development (Week 6–7; 10–12 hours)

Task	Duration	Deliverable
Decision Tree model (with hyperparameter tuning)	2 hrs	DT model; feature importance plot
Random Forest model (200 trees; parameter tuning)	3 hrs	RF model; feature importance ranking
Gradient Boosting model (XGBoost; tuning)	3 hrs	GBM model; feature importance
SVR model (kernel tuning; scaling applied)	2 hrs	SVR model; performance metrics
Model comparison table; select primary model	2 hrs	Model comparison scorecard

Expected Outcome:

- Random Forest: $R^2 \approx 0.78–0.85$, RMSE ≈ 1.8–2.5 days
- Gradient Boosting: $R^2 \approx 0.80–0.88$, RMSE ≈ 1.5–2.2 days

---

Phase 5: Validation & Hypothesis Testing (Week 7–8; 5–6 hours)

Objective: Validate models; test hypotheses rigorously.

Task	Duration	Deliverable
Time-series cross-validation (rolling origin)	2 hrs	CV results across time splits
Feature importance analysis; map to hypotheses	1.5 hrs	Feature-hypothesis mapping
Residual analysis; model assumptions testing	1 hr	Residual plots; assumption tests
Prediction interval analysis (uncertainty quantification)	0.5 hrs	Interval coverage; calibration plot
Final hypothesis scorecard; findings summary	1 hr	Comprehensive hypothesis validation report

Expected Outcome:

- ≥5 hypotheses strongly supported ($p<0.05$)
- Model assumptions reasonably met
- Prediction intervals well-calibrated

Phase 6: Documentation & Finalization (Week 8; 2–3 hours)

Objective: Prepare comprehensive documentation.

Task	Duration	Deliverable
Compile comprehensive modeling report	1 hr	Full modeling strategy report
Create visualizations: model performance, predictions	0.5 hrs	Performance plots; example predictions
Executive summary slide deck; key findings	1 hr	5–8 slide presentation for stakeholders
Code repository organization; README	0.5 hrs	GitHub repo; reproducible code

---

5.2 Key Deliverables Checklist

Deliverable 1: Problem Definition & Hypothesis Report

- [ ] Problem statement: Shelf-life prediction objective & business value
- [ ] 6 formulated hypotheses with mechanistic reasoning
- [ ] EDA validation for each hypothesis (plots, p-values)
- [ ] Supported hypotheses scorecard

Deliverable 2: Model Development Report

- [ ] Feature selection methodology & rationale
- [ ] 4 trained models (Linear, DT, RF, GBM) with configurations
- [ ] Model performance comparison table
- [ ] Primary model justification (Random Forest)

Deliverable 3: Evaluation & Validation Report

- [ ] Test set performance: R^2, RMSE, MAE, MAPE, MedAE
- [ ] Cross-validation results (5-fold + time-series folds)
- [ ] Feature importance rankings (mapped to hypotheses)
- [ ] Residual analysis; model assumption testing

Deliverable 4: Hypothesis Testing Report

- [ ] Regression coefficients with p-values (H1–H6)
- [ ] Feature importance validation (H1–H6 mapping)
- [ ] Mechanistic alignment (model findings ↔ food science theory)
- [ ] Final hypothesis scorecard

Deliverable 5: Code Repository

- [ ] `problem_definition.py` — Problem setup
- [ ] `data_preparation.py` — Cleaning, transformation, splitting
- [ ] `model_training.py` — All model implementations
- [ ] `model_evaluation.py` — Validation, metrics, visualization
- [ ] `hypothesis_testing.py` — Statistical tests, hypothesis mapping
- [ ] `README.md` — Usage guide
- [ ] `requirements.txt` — Package dependencies

Deliverable 6: Stakeholder Communication

- [ ] Executive summary (2–3 pages)
- [ ] Slide deck (5–8 slides)

EXPECTED OUTCOMES & SUCCESS CRITERIA

6.1 Model Performance Success Criteria

Criterion	Target	Rationale
R^2 (Test Set)	≥ 0.78	Explains ≥78% of shelf-life variance
RMSE	≤ 2.2 days	Acceptable error for business decisions
MAE	≤ 1.8 days	Median error magnitude acceptable
MAPE	≤ 13%	Percentage error acceptable
CV Stability	CV std(R^2) ≤ 0.04	Model generalizes well
Generalization	Test R^2 - Train R^2 ≤ 0.05	Low overfitting risk
Prediction Intervals	95% CI coverage ≈ 0.93–0.97	Uncertainty quantification accurate

6.2 Hypothesis Testing Success Criteria

Criterion	Target	Rationale
Supported Hypotheses	≥ 5 out of 6	Most hypotheses validated
Feature Significance	≥ 5 features with $p < 0.05$	Statistical validation
Effect Size Match	Coefficient direction matches expectation	Predictions align with food science
Feature Importance	Hypothesis-related features in top 8	Hypotheses capture key predictors
Residual Normality	Shapiro-Wilk $p > 0.05$	Model assumptions met

6.3 Business Impact Success Criteria

Criterion	Target	Benefit
Waste Reduction	Model reduces premature rejections by 20%	Cost savings: 5–10% of COGS
Food Safety	Model flags 95%+ at-risk batches	Recall prevention
Decision Support	Used in ≥50% of batch decisions within 6 months	Operational adoption
Stakeholder Satisfaction	Rating ≥ 4/5 for clarity and utility	Project success

CONCLUSION

This comprehensive predictive modeling strategy bridges the feature engineering work of Week 3 with practical, hypothesis-driven modeling. By:

1. **Clearly defining** the shelf-life prediction problem
2. **Formulating 6 mechanistically grounded hypotheses**
3. **Selecting multiple model types** (linear, tree-based, ensemble)
4. **Establishing rigorous evaluation metrics** (R^2, RMSE, MAE, cross-validation)
5. **Planning a phased, 4-week implementation** with clear milestones

Organizations can build predictive models that are not only accurate but also scientifically defensible and actionable.

REFERENCES

[1] Rashvand, M., et al. (2025). Artificial intelligence for prediction of shelf-life of various food products: A review. *Food Control*, 156, 110201. https://doi.org/10.1016/j.foodcont.2025.110201 — AI/ML for shelf-life prediction; hybrid models, non-invasive testing.

[2] Tarlak, F., et al. (2023). The use of predictive microbiology for shelf-life estimation and safety assessment. *Food Microbiology*, 112, 104224. — Predictive microbiology models; machine learning applications.

[3] Singh, R. P., & Heldman, D. R. (2014). *Introduction to food engineering* (5th ed.). Academic Press. — Arrhenius kinetics, thermal processing, shelf-life models.

[4] Kuhn, M., & Johnson, K. (2020). *Feature engineering and selection: Practical approaches for predictive models* (2nd ed.). Chapman & Hall/CRC Press. — Feature selection, model validation, hypothesis testing.

[5] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. — Foundational random forest methodology; feature importance.

Document Prepared By: Data Science Team
Last Updated: December 2025
Classification: Internal Use
Version: 1.0

WEEK 4 TASK: COMPLETE

 5 Sections Delivered:

1. ✅ **Problem Definition** — Shelf-life prediction with business rationale
2. ✅ **Hypothesis Formulation** — 6 mechanistic hypotheses with testing strategy
3. ✅ **Model Selection** — 5 models with pros/cons; Random Forest recommended
4. ✅ **Evaluation Metrics** — RMSE, MAE, R^2, MAPE, MedAE, cross-validation framework
5. ✅ **Implementation Roadmap** — 6-phase 30–35 hour plan with deliverables

NO CODING. NO DATA. NO MODELS TRAINED. Pure strategic planning documentation as required.