Predictive Modeling for Dairy Feed Optimization and Milk Yield Prediction: A Machine Learning Approach

Md Rashed mosharof
Shanghai University of Engineering Science

Abstract

The relationship between feed composition and milk yield in dairy cows is of paramount importance for optimizing dairy farm productivity. This paper presents a machine learning-based framework for predicting daily milk yield based on various factors, such as feed composition, cow characteristics, and environmental conditions. Multiple machine learning models—Linear Regression, Random Forest, Gradient Boosting, and Feedforward Neural Networks—were implemented and compared based on their performance in predicting milk yield. The study demonstrates that Linear Regression achieves the best balance between accuracy and interpretability, while Random Forest and Gradient Boosting models exhibit enhanced robustness to non-linearities in the dataset. This paper provides a detailed explanation of the models used, their mathematical formulations, and future research directions to further refine precision agriculture practices.

1 Introduction

In the dairy industry, milk production is influenced by several factors, such as feed composition, environmental conditions, and cow-specific characteristics (e.g., breed, weight, and age). Farmers are often faced with the challenge of balancing feed costs with milk yield. Traditionally, optimizing this balance has been approached empirically, but recent advancements in data science and machine learning offer new possibilities for precision dairy farming.

The goal of this project is to design predictive models that can accurately estimate daily milk yield based on the provided features. Using machine learning models, we aim to predict milk yield efficiently and provide farmers with actionable insights into how different feed compositions and environmental factors affect milk production. The models can also help in optimizing feed costs by predicting the milk yield for various feed compositions.

2 Dataset Description

2.1 Feature Breakdown

The dataset used in this study is a synthetic dataset representing real-world dairy farming conditions. It consists of the following key features:

- **Cow Breed**: Different breeds have varying genetic potential for milk production.
- **Cow Age (years)**: Age influences milk yield, with older cows often producing less milk than younger ones in their prime milking years.
- **Cow Weight (kg)**: Weight affects a cow's metabolism and milk production efficiency.
- **Lactation Stage**: The cow's lactation stage significantly impacts the quantity of milk produced.
- **Daily Feed (kg)**: Total daily feed intake in kilograms.
- **Feed Composition**: Percentages of protein, fat, and fiber in the feed, which directly influence milk production.
- **Environmental Factors**: Daily temperature and humidity, which can impact a cow's metabolic efficiency and, consequently, its milk production.

2.2 Target Variable

The target variable for the machine learning models is **Daily Milk Yield (liters/day)**, which is the total amount of milk produced by each cow every day.

2.3 Data Preprocessing

Before feeding the data into machine learning models, preprocessing steps were applied:

- **Handling Missing Data**: Synthetic data generally does not contain missing values. In real-world scenarios, techniques such as imputation could be applied.
- **One-Hot Encoding**: The categorical variables, such as breed and lactation stage, were one-hot encoded.
- **Feature Scaling**: All numerical variables were scaled using Z-score normalization:

  $$Z = \frac{X - \mu}{\sigma}$$

  where Z is the normalized value, X is the original value, μ is the mean of the feature, and σ is the standard deviation.

3 Model Descriptions

Four machine learning models were employed to predict daily milk yield based on the given features.

3.1 Linear Regression

Linear Regression assumes a linear relationship between the features and the target variable:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon$$

Where Y is the predicted milk yield, X_i are the input features, β_i are the coefficients, and ϵ is the error term. Linear Regression is trained by minimizing the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

3.2 Random Forest Regressor

Random Forest is an ensemble learning method that creates a collection of decision trees. The prediction is the average of the predictions made by the individual trees:

$$\hat{Y} = \frac{1}{T} \sum_{t=1}^{T} h_t(X)$$

Where T is the total number of trees, and $h_t(X)$ is the prediction made by the t-th tree.

3.3 Gradient Boosting Regressor

Gradient Boosting builds trees sequentially, with each tree correcting the errors of the previous one:

$$\hat{Y}_t = \hat{Y}_{t-1} + \eta \cdot h_t(X)$$

where η is the learning rate.

3.4 Feedforward Neural Network (FNN)

A simple FNN with one hidden layer can be expressed as:

$$\hat{Y} = f(W_2 \cdot f(W_1 \cdot X + b_1) + b_2)$$

Where W_1, W_2 are weight matrices, b_1, b_2 are bias terms, and f is the activation function (ReLU in this case).

4 Model Performance

The models were evaluated using Mean Absolute Error (MAE) and R^2 Score.

Model	**MAE**	**R^2**
Linear Regression	0.1189	0.9988
Random Forest	0.3187	0.9927
Gradient Boosting	0.1882	0.9975
Neural Network	0.4560	0.9826

Table 1: Model Performance Comparison

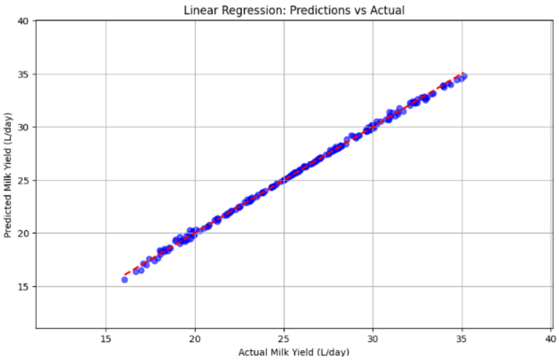

Figure 1: Linear Regression model performance.

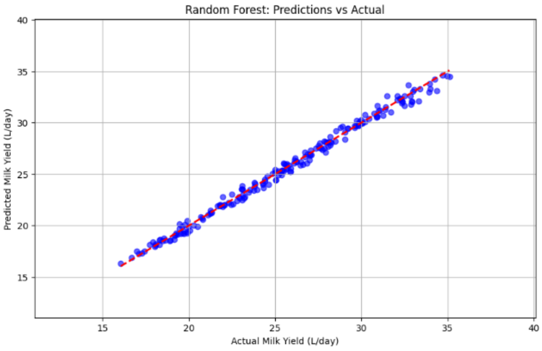

Figure 2: Random Forest model performance.

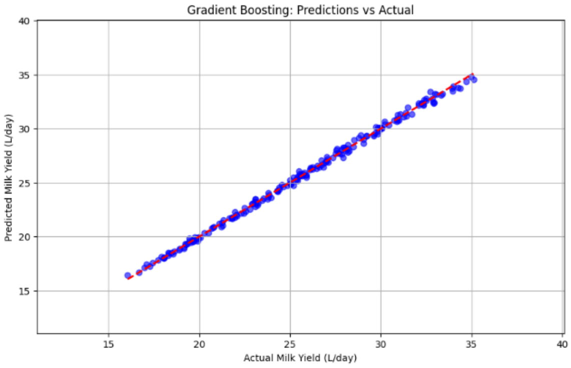

Figure 3: Gradient Boosting model performance.

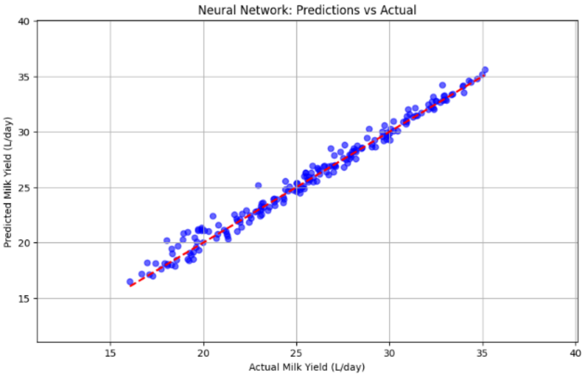

Figure 4: Neural Network model performance.

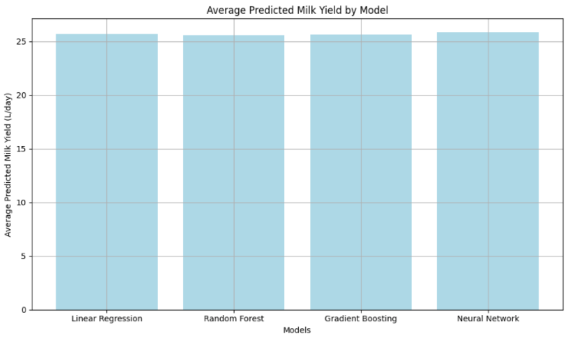

Figure 5: Average milk yield across different models.

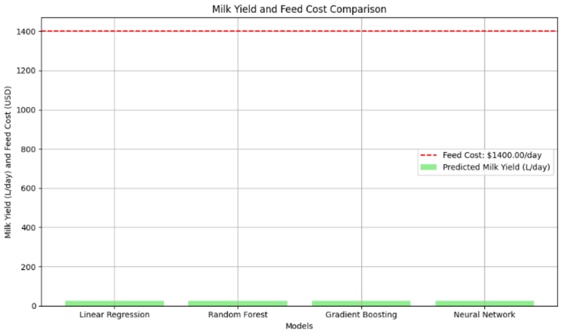

Figure 6: Comparison of milk yield and feed costs.

5 Future Work

Future work could include:

- **Data Collection from Real-World Farms**: Incorporate real-world data to improve model robustness.
- **Temporal Models**: Use time-dependent models like RNNs or LSTMs.
- **Feed Cost Optimization**: Optimize both milk yield and feed costs.
- **IoT Integration**: Use real-time data from IoT devices for dynamic predictions.
- **Sustainability Factors**: Explore ways to reduce environmental impact.

6 Conclusion

This project demonstrates the utility of machine learning models in predicting dairy milk yield based on cow characteristics, feed composition, and environmental conditions. Linear Regression proved to be the most effective, though more complex models like Random Forest and Gradient Boosting capture non-linear interactions well. Future work will focus on real-world data, temporal dynamics, and sustainability.