Capstone Project 2 – Final Report

Identifying Key Drivers of Revenue in Dairy Goods Sales

Wen Yang

Problem Statement

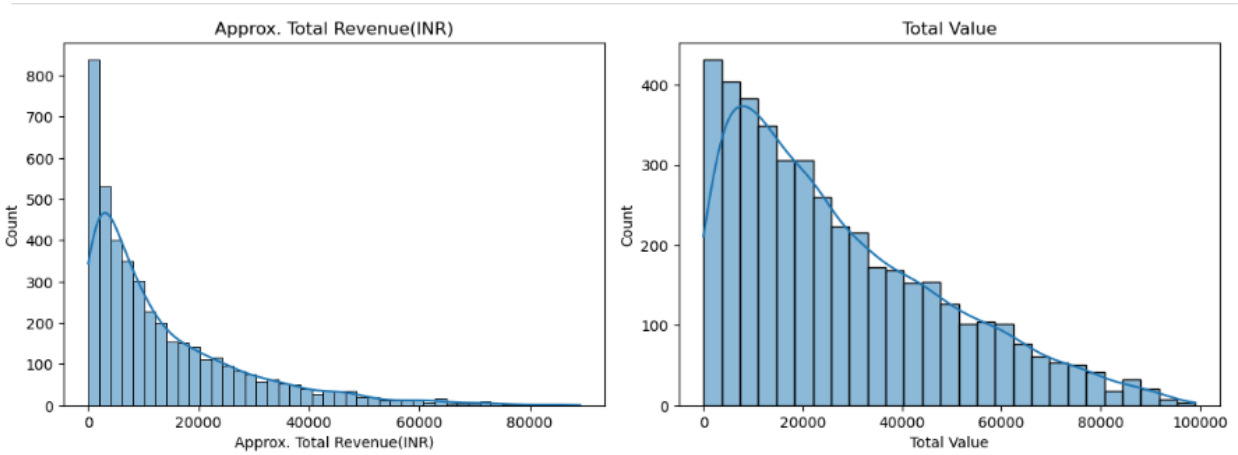
The dairy industry faces ongoing challenges balancing product freshness, pricing strategy, inventory levels, and production capacity. Because dairy goods are highly perishable, producers must carefully coordinate production, storage, and distribution decisions to avoid spoilage while meeting fluctuating customer demand. Although the Dairy Goods Sales dataset (2019–2022) contains detailed information on farm operations, product characteristics, shelf life, storage conditions, and sales activity, dairy producers often lack a systematic, data-driven understanding of which factors most strongly influence total revenue.

This project addresses the question: how can dairy producers leverage historical sales and operational data to identify the key factors influencing total revenue, and use these insights to improve pricing and operational decision-making?

Data Wrangling and Exploratory Data Analysis

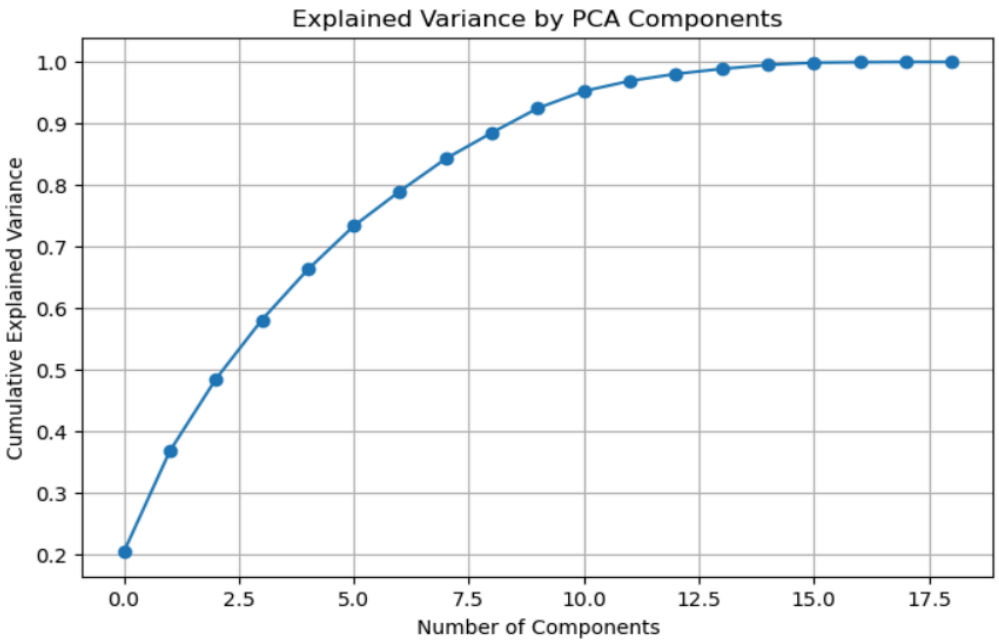
The exploratory phase involved understanding the structure and behavior of 4,325 dairy sales records spanning multiple product categories, brands, customer locations, and storage types. After validating date fields and ensuring internal consistency, I examined the dataset for duplicates, missing values, and impossible values. No missing values were found.

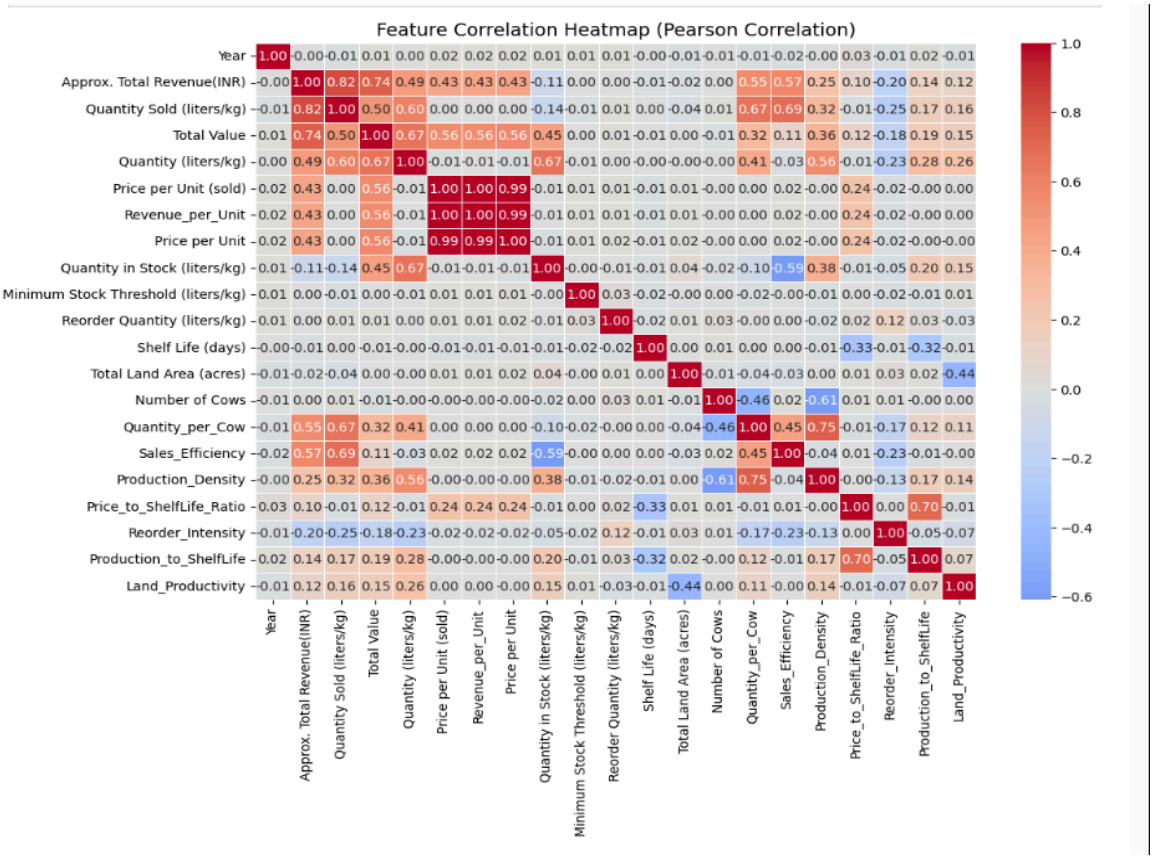
I selected Approx. Total Revenue (INR) as the target because it best represented true business outcomes. Features were grouped into numerical, categorical, and date-derived categories. Additional engineered variables were created to reflect essential operational behaviors, such as Quantity_per_Cow, Sales_Efficiency, Production_Density, Price_to_ShelfLife_Ratio, Land_Productivity, Reorder_to_Stock_Ratio, and Reorder_Intensity.

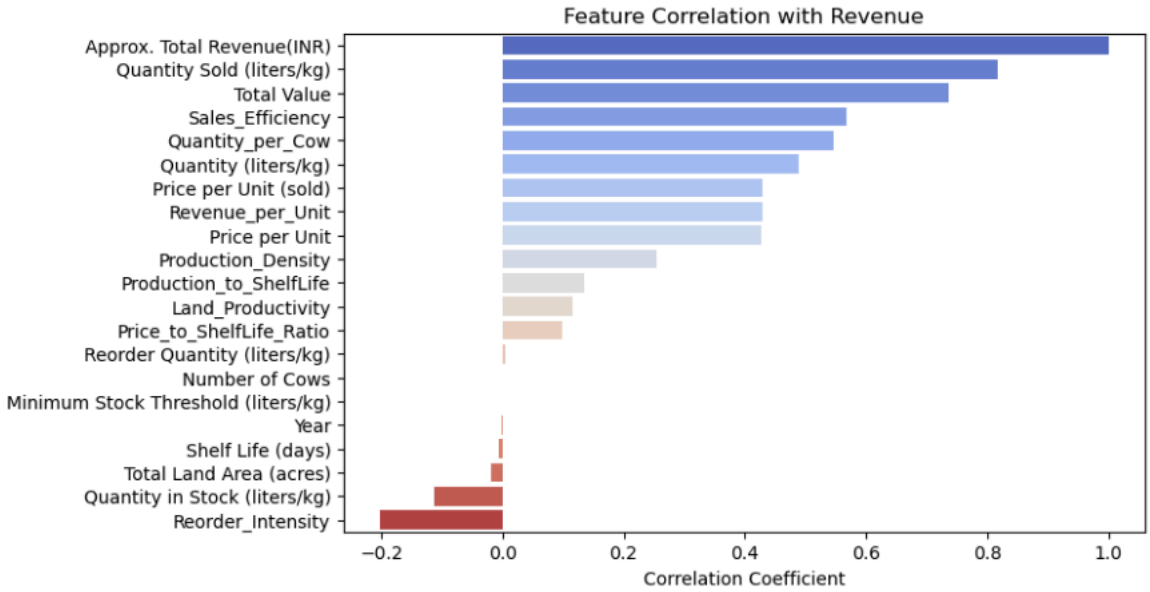


Univariate distributions revealed right-skewed revenue and quantity figures, with several high-volume transactions dominating the revenue landscape. Bivariate analysis showed that operational efficiency metrics, rather than brand or sales channel, had the strongest explanatory relationship with revenue. ANOVA confirmed statistically significant revenue differences across product categories and storage conditions, while brand and sales channel did not demonstrate meaningful variation.

Correlation analysis indicated strong dependencies among revenue-formula components such as Quantity Sold, Total Value, and Price per Unit. To prevent data leakage, these mathematically dependent features were removed. Remaining numerical features exhibited acceptable correlation levels. PCA revealed that the first ten principal components explained approximately 95% of total variance, with operational efficiency metrics dominating the feature space. Taken together, these exploratory findings demonstrated that operational production dynamics—not branding or distribution—primarily determine revenue outcomes. The dataset was cleaned, engineered, and validated for predictive modeling.


Explained Variance by PCA Components
Cumulative Explained Variance
Number of Components
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Feature Correlation Heatmap (Pearson Correlation)
Year
Approx. Total Revenue(INR)
Quantity Sold (liters/kg)
Total Value
Quantity (liters/kg)
Price per Unit (sold)
Revenue_per_Unit
Price per Unit
Quantity in Stock (liters/kg)
Minimum Stock Threshold (liters/kg)
Reorder Quantity (liters/kg)
Shelf Life (days)
Total Land Area (acres)
Number of Cows
Quantity_per_Cow
Sales_Efficiency
Production_Density
Price_to_ShelfLife_Ratio
Reorder_Intensity
Production_to_ShelfLife
Land_Productivity
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6



Feature Engineering

Feature engineering served as a critical bridge between raw operational data and business-meaningful predictive insights. Several new metrics were constructed to reflect efficiency, shelf-life utilization, production intensity, pricing strategy, and demand behavior.

> Quantity_per_Cow: a measure of per-animal productivity—emerged as the single most powerful indicator of revenue potential;
> Sales_Efficiency: captured how effectively production translated into actual sales;
> Production_Density: reflected output intensity relative to herd size;
> Price_to_ShelfLife_Ratio: quantified whether products were priced appropriately relative to perishability;
> Reorder_Intensity and related stock-turn ratios: represented Demand strength.

Together these engineered features transformed the dataset from a collection of raw inputs into a structured set of interpretable business metrics, enabling the models to learn more nuanced relationships.

Preprocessing and Training Data Development

Categorical variables (Product, Brand, Storage Condition, Quarter, Customer Location, etc.) were one-hot encoded using drop_first=True to reduce multicollinearity. Numeric features

were standardized using StandardScaler to ensure equal contribution during model training.

Final dataset dimensions:

- 59 input features
- 1 target: Approx. Total Revenue (INR)
- 4,037 rows

An 80/20 train–test split was applied (3,229 training rows; 808 testing rows). Preprocessed datasets were saved as PKL files to preserve scaling and encoding structures, which CSV cannot retain.

Model Development and Evaluation

Three supervised machine learning models were evaluated: Linear Regression, Random Forest, and Gradient Boosting. The objective was not forecasting accuracy but constructing a sufficiently strong model to reliably expose underlying feature influence.

Linear Regression underfit the nonlinear operational patterns, showing the weakest performance. Random Forest performed substantially better, capturing interactions and nonlinear behavior with low generalization error. However, Gradient Boosting—after hyperparameter tuning (n_estimators=300, learning_rate=0.1, max_depth=4)—produced the best learning behavior and the strongest stability.

Model Performance:

Linear Regression:

- CV MAE ≈ 7610.17 ± 353.07
- Test MAE = 7353.60
- Test R^2 = 0.527

Random Forest:

- CV MAE ≈ 3,632.37 ± 168.24
- Test MAE: 3,495.11
- Test R^2: 0.866

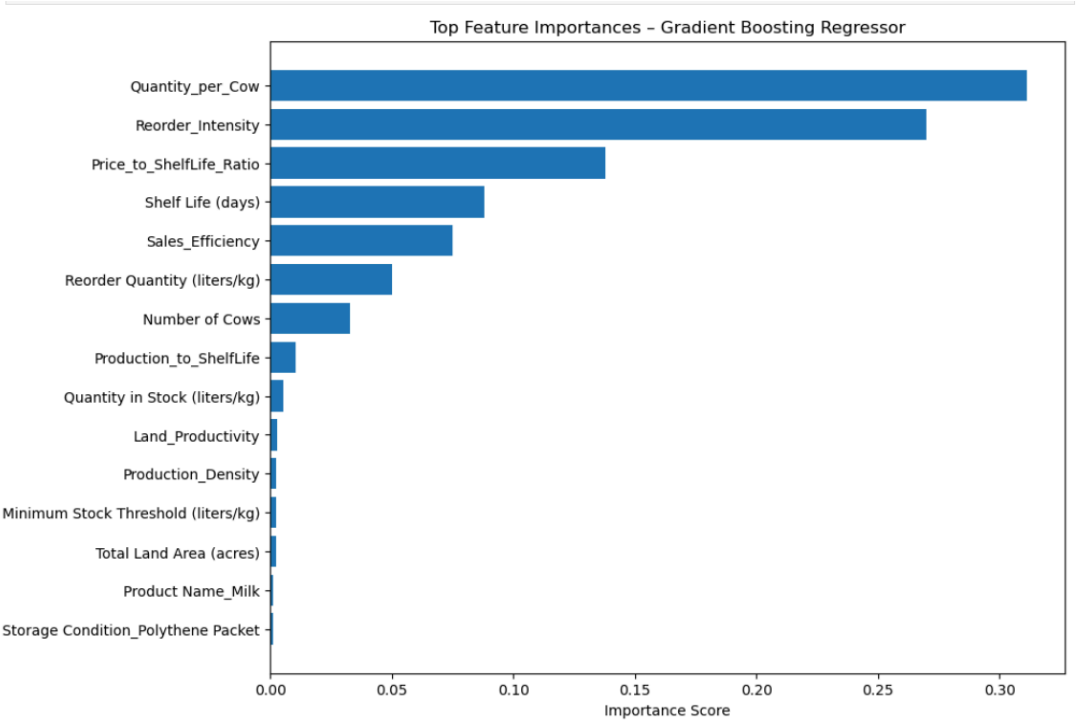
Gradient Boosting (Final Model):

- CV MAE ≈ 2588.87 ± 114.13
- Test MAE: 2538.40
- Test R^2: 0.924

Although Gradient Boosting provides strong predictive accuracy, its primary value in this project is interpretive: it most effectively captures complex operational relationships, enabling a reliable extraction of revenue drivers.

Key Drivers of Revenue:

The Gradient Boosting importance ranking revealed a consistent and interpretable hierarchy of operational influences.



Quantity_per_Cow emerged as the strongest driver of revenue. This metric reflects per-animal productivity, indicating that improvements in herd efficiency translate directly into financial gains.

Reorder_Intensity was the second most influential variable. High reorder rates represent strong, sustained product demand and serve as a signal of market pull and product–market fit.

Price_to_ShelfLife_Ratio ranked third, showing that perishability-adjusted pricing plays a critical role in revenue performance. Products priced appropriately relative to shelf life generate stronger, more stable returns.

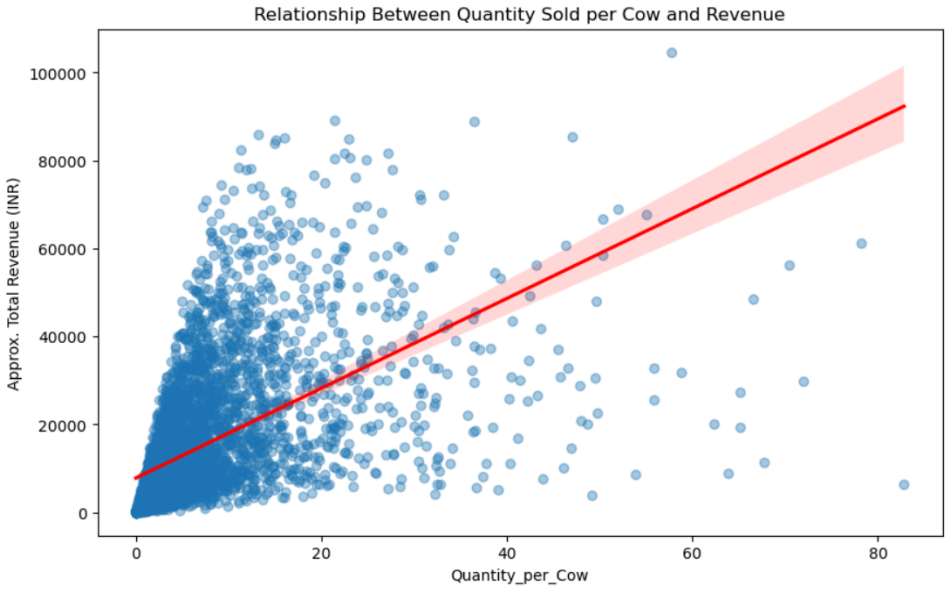
Other important features included Sales_Efficiency, Shelf Life, Reorder Quantity, Number of Cows, and Land Productivity. Together they form a coherent operational narrative: revenue in dairy operations is determined primarily by production efficiency, demand strength, resource utilization, and perishability-aligned pricing.

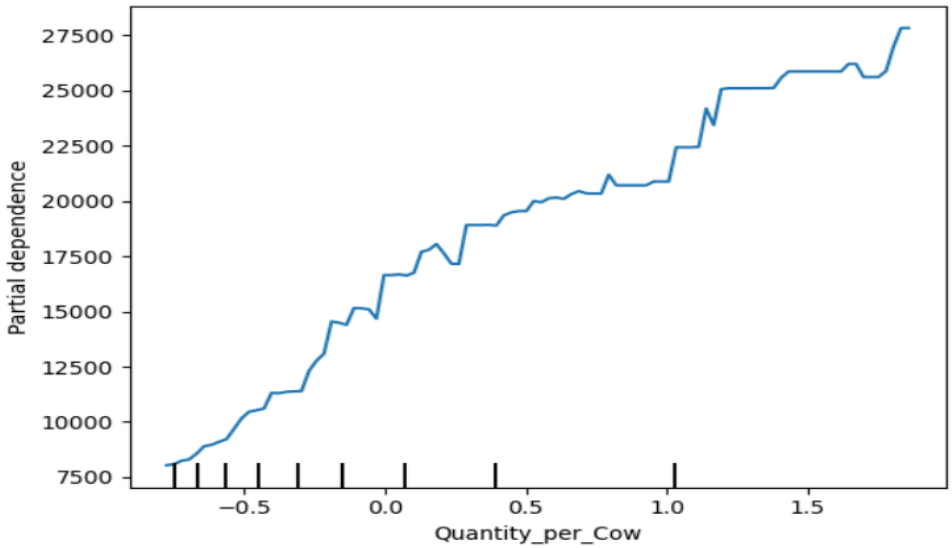
Importantly, the model was not used to generate future revenue predictions. Instead, it functioned as a diagnostic tool to uncover the structural drivers of revenue and quantify their relative influence.

Recommendations

Three actionable recommendations arise directly from the model insights:

1. Improve per-cow productivity: Since Quantity_per_Cow is the strongest driver, optimizing feed strategy, herd health, and production processes can yield substantial revenue improvements.





2. Track and leverage Reorder_Intensity: Consistently high reorder behavior signals strong demand; such products should be prioritized for distribution, production scaling, or expansion into new markets.

3. Align pricing strategies with perishability: Adjusting Price_to_ShelfLife_Ratio can help maximize revenue while reducing wastage. Premium pricing may be justified for products that maintain strong sales despite short shelf lives.

Limitations

The dataset does not include external influences such as climate variation, competitive pricing, or supply chain disruptions. Cost data is absent, meaning that the analysis focuses on revenue rather than profit. Some engineered variables may still exhibit correlated structures that influence interpretability.

Future Research

Future work may incorporate time-series analysis for seasonality detection, cost modeling for profit optimization, or SHAP interaction studies to quantify nonlinear dependencies between production, shelf life, and demand intensity. Integrating macroeconomic or environmental variables could strengthen real-world applicability.

Conclusion

This project successfully identified the key operational and commercial drivers of Approx. Total Revenue in dairy goods. Gradient Boosting offered the strongest interpretive foundation, enabling the extraction of meaningful, business-aligned insights. The dominant drivers—Quantity_per_Cow, Reorder_Intensity, and Price_to_ShelfLife_Ratio—highlight the importance of production efficiency, demand behavior, and perishability-aware pricing in shaping financial performance.