Progress Report

Project Overview

This project aims to explore the underutilized data generated by dairy production systems, specifically from cows. Often, this data is not well integrated with other sources, yet it offers valuable insights into the relationships between production output and various conditions. Utilizing the Lely Horizon application, we will download the necessary files for comprehensive analysis of these datasets. By leveraging farm management applications, we can forecast performance and optimize farming practices based on the insights gained.

Key Metrics

We are interested in **milk production** and **milk quality**, and we are trying to find which other variables are correlated with them.

Terminology

Some words are highlighted as high frequency in this project, as it is based on agriculture.

1. **Lactation:** Lactation is the period when mammals produce milk after giving birth to offspring. It is part of the reproductive cycle, aimed at feeding newborns.
2. **Continuous milking**: Continuous milking refers to the process of extracting milk from the mammary glands throughout the entire lactation period, including the usual dry period.
3. **Dry period:** The dry period is the stage when dairy cows do not produce milk between two lactation cycles, typically used to restore mammary gland health.

Literature Review

1. **Integrating Big Data in Dairy Farm Decision-Making**

The evolution of data-driven decision-making in dairy farming has seen significant strides with the integration of big data analytics, precision farming, and Internet of Things (IoT) technologies. As discussed by Cabrera et al. (2020), the development of the Dairy Brain initiative showcases a real-time, continuous decision-making engine that leverages vast data

streams from dairy farms. This system aims to optimize various aspects of dairy farm management, including nutritional grouping and health monitoring, thereby enhancing overall farm efficiency and productivity.

Similarly, Ferris et al. (2020) introduce the Dairy Brain decision support system, which integrates data across multiple farm management areas to provide actionable insights to farmers. This system utilizes a complex array of data inputs from cow and herd health to economic indicators, employing machine learning and optimization techniques to forecast and prescribe management strategies. The integration of such data not only aims at improving immediate farm operations but also addresses broader economic and environmental impacts.

2. Challenges and Innovations in Data Utilization

Despite the availability of extensive data streams, the challenge remains in the effective integration and real-time application of this data to improve decision-making on dairy farms. Cabrera et al. (2020) highlight the practical applications of their system, such as the detection of clinical mastitis and nutritional management, which are made possible by continuous and dynamic data analysis. These applications demonstrate the potential to reduce operational costs and enhance animal health management, which are critical for the sustainability and profitability of dairy operations.

Ferris et al. (2020) also discuss the operationalization of data through the Agricultural Data Hub (AgDH), which facilitates the aggregation and analysis of data from various sources. This approach not only enhances the decision-making process but also ensures the adaptability of the system to the varying needs of different dairy farms. The flexibility and scalability offered by such systems are essential for the future of precision dairy farming.

3. Implications for Dairy Farm Management

The integration of big data into dairy farming practices presents a transformative potential for the industry. The ability to make informed decisions based on comprehensive data analysis can significantly impact the efficiency and sustainability of farm operations. For instance, the predictive analytics used in monitoring herd health can preemptively address potential issues, reducing the incidence of disease and associated costs.

Moreover, the case studies presented in the literature exemplify the practical benefits of these systems in real-world settings. These include improvements in feed efficiency, reduction in disease incidence, and optimization of milk production through tailored animal management strategies. These advancements underscore the critical role of technology and data in shaping the future of dairy farming.

4. **Conclusion**

The review of the literature reveals a significant shift towards data-driven decision-making in dairy farming. The development and implementation of systems like Dairy Brain represent a major step forward in optimizing farm management through the integration of big data. The ongoing innovations and improvements in data analytics and machine learning are likely to continue driving efficiency and sustainability in the dairy industry, presenting a promising outlook for the application of these technologies in agricultural practices.

Data Analysis

In our data analysis, we set Day production, the topic of greatest concern to the host, as the target variable and conducted in-depth exploration of the dataset. We selected 'Lactation days', 'Number of Calves', and 'Number of Lactations' as potential primary factors influencing Day production and analyzed them using linear modeling.

Our analysis revealed that the overall model is significant (F-statistic p-value < 2.2e-16). Specifically, 'Lactation days' and 'Number of Lactations' have significant effects on 'Day production', while the impact of 'Number of Calves' is relatively minor. The model's R-squared value is approximately 0.4481, indicating that the model explains around 44.81% of the variability in the target variable, but we believe this is still insufficient.

To further investigate model performance, we identified three potential issues:

1. Bias in the dataset: Missing values in the dataset may affect model performance as they could lead to inaccurate estimations of real-world scenarios. Therefore, we plan to conduct more thorough data cleaning and preprocessing, including handling missing values, addressing outliers, and data transformation, to ensure data quality and reliability.

2. Inappropriateness of linear regression: Considering that real-world relationships may be more complex, we intend to explore alternative models such as polynomial regression, decision trees, and random forests to better capture potential nonlinear relationships and comprehensively evaluate and compare model performance.
3. Inadequate variable selection: We plan to refine variable selection through feature engineering and screening to choose variables with higher predictive power or experiment with different variable combinations to enhance model performance. Additionally, we will employ techniques like cross-validation to assess model generalization ability, ensuring model reliability and robustness.

Through this iterative process, we aim to develop a more accurate and interpretable model to address the host's concerns.

Challenges and Solutions

We face several challenges in our project. Firstly, while we can access datasets through Lely Horizon, they are often small and sometimes even empty, likely due to lack of cleaning. To address this, we initially conduct basic analyses on these datasets to identify trends or insights. We plan to request larger, more comprehensive datasets from our host for more robust analysis. Secondly, handling missing data poses a challenge. Currently, we opt to delete rows with missing values, but we are unsure if this is the best approach. We intend to consult with our host to explore alternative methods for handling missing data effectively. Moreover, while we have identified variables of interest, refining our research questions requires further discussion with our host to ensure specificity and focus in our analysis. Lastly, we possess literature review papers that could guide our methodology. We must decide whether to adopt a method outlined in one of these papers or explore different analysis models from various sources to make informed comparisons.

Reference

Ferris, M. C., Christensen, A., & Wangen, S. R. (2020). Symposium review: Dairy Brain—Informing decisions on dairy farms using data analytics. *Journal of Dairy Science, 103(10),* 3874–3881. https://doi.org/10.3168/jds.2019-17199

Madsen, T. G., Nielsen, M. O., Andersen, J. B., & Ingvartsen, K. L. (2008). Continuous lactation in dairy cows: Effect on milk production and mammary nutrient supply and extraction. *Journal of Dairy Science, 91(7),* 1791–1801. https://doi.org/10.3168/jds.2007-0905