Making use of dairy herd improvement records and machine learning to identify best management strategies

D. Warner[1], J. Gunn[1] and D.E. Santschi[1]

Lactanet, 555, boul. des Anciens-Combattants, Sainte-Anne-de-Bellevue, H9X 3R4, Québec, Canada
Corresponding Author: dwarner@lactanet.ca

Abstract

Dairy herds continuously generate on-farm data that have a high potential for decision making support. However, variability among herds can be high and it can be, therefore, challenging for dairy producers and extension services to interpret farm records and identify best management practices. In this observational study, we used DHI data to identify the main factors explaining the overall herd performance for 71 commercial Holstein herds in New Brunswick, Canada. A machine learning approach was used to generate high-level recommendations as well as detailed herd-specific recommendations. Enrolled farms were equipped with a parlour (n = 31), pipeline (24) or automatic (16) milking system. Herd performance records from September to December 2020 were assembled from a DHI database and included 71 indicators related to production, reproduction, health, and longevity. Herds were segmented based on their overall herd performance using hierarchical clustering based on principal components. The principal components were used primarily as pre-processing step to de-noise the data and to balance the influence of the similar DHI records. In-dept analysis of clusters was conducted using decision tree induction with the aim to generate an interpretable on-farm decision making tool. Three herd clusters were identified and consisted of low overall herd performance across all DHI indicators (Cluster 3; n = 17), a medium overall herd performance but high longevity (Cluster 1; n = 36), and high overall herd performance but with a concomitant low longevity (Cluster 2; n = 18). Decision tree induction further allowed to identify the most important DHI indicators explaining the adherence of a herd to one of the three clusters. This ultimately allowed to establish high-level recommendations and visualize which performance indicators a herd might want to focus on to improve overall herd performance. For instance, low performing herds were observed to be mainly driven by a low reproduction performance. In addition, single predictions of the decision tree algorithm were fitted with a local interpretable model. This latter approach allowed to add interpretability to the decision tree model and generate dynamic herd-specific recommendations for each farm. In conclusion, the results suggest that mining DHI data can give a valuable insight into best herd management practices and can be used to highlight opportunities for improvements.

Keywords: Decision support, DHI, artificial intelligence, decision tree.

Introduction

Dairy herds continuously generate on-farm data that have a high potential for decision making support. However, variability among herds can be high and it can be, therefore, challenging for dairy producers to interpret farm records and identify best management practices. In addition, on-farm data from dairy herd improvement (DHI) records are typically complex, incomplete, variable over time, and correlated to each other. With such datasets, traditional inferential statistical methods often perform poorly (Heald *et al*., 2000). A machine learning approach might be more suitable for DHI data as these types of algorithms provide more flexibility with incomplete data and generally handle better interactions among predictors and non-linear relationships (Abbot, 2014). Unsupervised machine learning algorithms such as clustering are typically used to find hidden patterns in complex and large datasets, whereas supervised machine learning algorithms are typically used for generalized predictions. In addition, machine learning is particularly suitable for field applications as it can be implemented and automated for field applications (Dominiak and Kristensen, 2017). As such, a combined unsupervised and supervised machine learning approach was used in the present study to find groupings in data originating from the dairy herd population in New Brunswick, Canada, and to identify management practices of high performing dairy herds as opposed to other dairy herds for decision making support to improve herd performance and profitability.

Material and methods

Participating herds

A total of 89 dairy herds across New Brunswick for which DHI records are routinely collected were initially included in the study. Only herds predominantly composed of Holstein cows (71 herds) were considered for further analysis as herd management was likely influenced to a large degree by the breed. Among these, 31 herds had a parlour, 24 herds had a pipeline, and 16 herds had an automatic milking system. Each participant received free access to the Lactanet service for transition management (Transition Cow Index; Nordlund, 2006) and subclinical ketosis screening (Foss, 2009) during the entire length of the study period which were used as input data in the prediction model.

Data collection

Dairy herd improvement data.

Test day records were collected and included 12-month average data on herd demography, production, longevity, and reproduction based on the last test records registered between September and December 2020. Data was cleaned to remove implausible values and outliers by removing the 1st and 99th percentile. Originally 71 variables were considered for data analysis. Variables missing at random were imputed using a random forest approach as described in van Buuren (2012), which generally handles well complex and inter correlated data in the presence of missing cases (Tang and Ishwaran, 2017) as it the case for DHI data. Data not missing at random (bodyweight, milk urea nitrogen records) were omitted from the analysis.

Survey data

Management practices were collected through an online survey administered by Lactanet technicians and advisors. The survey was previously developed and applied to collect data on approximately 2300 Quebec dairy herds with questions pertaining to the building, housing, feed bunk and feeds, bedding, cleaning, milking system, footbath, hoof trimming, exercise and pasture, calving, and drying off management (Lactanet, 2021). Among the 71 Holstein herds in New Brunswick, 42 completed the survey and were thus considered for analysis. Amongst these, 23 herds had a parlour, 22 herds a pipeline, and 12 herds an automatic milking system. The survey was completed between March 2020 and April 2021.

Data analysis

To identify the best management practices, herds were clustered into similar groups based on the overall herd performance using their respective test day DHI records by means of agglomerative hierarchical clustering on principal clustering (HCPC) using the FactoMinerR package in R (version 3.5.0; R Foundation for Statistical Computing, Vienna, Austria) of Le *et al.* (2018). Computation of the principal components was thereby done as pre-processing step to de-noise the data and to balance the influence of several groups of variables.

Differences in the DHI records among the identified clusters were first evaluated through a mixed effect linear regression with cluster considered fixed and milking type considered random using the lmerTest package (Kuznetsova *et al*., 2017).

The cluster output was investigated in more detail through a decision tree approach with the aim to predict the adherence to one of the three groups with the DHI indicators and identify the most meaningful indicators explaining the difference between the high, medium, and low performant herds. Decision tree induction was conducted through a CART classification tree as described in Breiman *et al*. (1984) using the caret modelling package workflow of Kuhn (2008). This approach allows to compute a tree-structured classification containing a collection of decision rules (represented by the tree branches) and cluster predictions (represented by the terminal node of a branch). Its simple and interpretable structure as well as its ability to compute decision rules in the presence of missing values using surrogate variables (i.e., substitutes for the primary splitter of a node when the primary splitter is missing; Rokach and Maimon, 2005) makes it an interesting tool for future field applications.

Prior to applying a decision tree algorithm, collinearity was checked to remove variables with a Spearman correlation coefficient above 0.60 as multicollinearity could lead to over-fitting, resulting in 19 final variables. The optimal model hyperparameters for the tree size (complexity parameter) were evaluated through a random search using a 5-fold cross-validation.

The global variable importance scores, that is the variables contributing most to the model development, were extracted based on the sum of the reduction in the loss function (information gain) attributed to each variable at each split. The model response was illustrated through an alluvial diagram for the variables with the higher global importance score. The variables were thereby normalized to a mean of 0 and a standard deviation of 1. Additional model agnostics based on the LIME approach (Local interpretable model-agnostic explanations; Ribeiro *et al*., 2016) were run to extract local variable importance scores, which allow to interpret the model outcome for each participating herd.

The survey data were analysed using a Fisher's test on the cluster output. The Fisher's test was preferred over the more commonly used chi-squared test due to the small sample number and failing to comply with the strict requirements of the latter (e.g., expecting a frequency never smaller than 1 and a frequency of 5 or more for at least 80% of the output cells). A list of promising management practices known to affect herd performance has been elaborated and presented to the participating producers. However, as the findings were statistically inconclusive due to the low response rate and large variability observed among the participating herds leading to an overall non-significant P-value (> 0.10), results from the survey are not shown here.

Each participant received a customize report with benchmarks for DHI records previously identified using decision tree induction as well as for some relevant management practices via a parameterized R markdown report (version 1.1.; Allaire *et al*., 2019) and were directly sent to the respective participant from the RStudio user interface via the utility package mailR (version 0.4.1; Premraj, 2015).

Results and discussion

Dairy herd improvement data

Three cluster were identified based on the herd DHI records and labelled as high, medium, and low overall herd performance. A detailed cluster description is shown in Table 1 for variables that differed significantly ($P \leq 0.05$) among clusters. The herd size did not differ significantly among the clusters ($P = 0.076$) but was somewhat higher for the Medium performance cluster due to a higher number of herds equipped with a milking parlour. The Low performance cluster did not contain any herds equipped with an automatic milking system. However, herds pertaining to the High performance cluster were not driven mainly by a specific milking system.

The cluster analysis suggested that the High performing herds were also the herds with the highest milk performance, milk value (mainly based on milk yields and components) and genetic potential. The somewhat higher genetic potential for energy-corrected milk for the Medium performant herds implies that some herds in this cluster do not exploit

Table 1. Cluster description for overall herd performance.

	Cluster		
Variable	High	Medium	Low
Number of herds	18	36	17
Herd size (median)	75.2	97.2	75.0
Milking system (number of herds)			
Automatic	6	10	0
Parlour	6	17	8
Pipeline (tie stall)	6	6	9
Production[1]			
kg ECM	10,996	9,909	8,468
kg fat	442	397	339
kg protein	361	328	277
Milk value (CAN$)	8,032	7,188	6,127
Genetic index for ECM	414	440	255
Reproduction and health[1]			
Transition Cow Index[2]	426	260	–173
Calving interval	405	403	451
Days to first breeding	82	81	100
Age at first calving	25.6	25	30
Days dry	67	63	71
% cows with SCC > 200,000	12.4	14.8	18.4
Longevity[1]			
% involuntary culled	29	16	23
% cows dead	4.7	2.8	5.3
% cows left	44	30	35
% cows left at 60 DIM	8.3	5.2	7.3
% cows left for reproduction	8.8	4.6	5
% cows left for feet problems	3.7	1.9	4.2

their genetic potential to the fullest as suggested by their lower milk performance and milk value. In addition, the High performing had the best transition cow management and a high reproduction performance. This further highlights the importance to pay attention to the non-producing cows and heifers within a herd. Nonetheless, these herds had a high involuntary culling rate and deficiencies in early lactation, in particular compared to the Medium performance cluster indicating a potential for improvement.

Using a decision tree approach, the most important variables were identified that explain the adherence of a herd to either the High, the Medium and Low performing herd cluster (Figure 2). The top performing herds could be predicted via a low calving interval, low age at first calving, high involuntary culling rate, high energy-corrected milk yield, high turnover rate and percentage of cows left, and an excellent transition cow management. The decision tree suggested that the contribution of the genetical potential was low overall suggesting that some lower performing herds might not fully exploit their genetic potential.

Using a model-agnostic approach, the model predictions were investigated in more detail. For a randomly selected herd (herd 61 in Figure 2), we noticed a high calving interval, suboptimal transition management and low turnover rate which were in line with other Low performing herds. However, its genetic potential was overall within the top 50% herds and thus considerably higher than that of its peers, suggesting that herd 61 does not fully exploit its high genetic potential. This approach allows therefore to investigate the decision rules for each individual herd and can help with decision

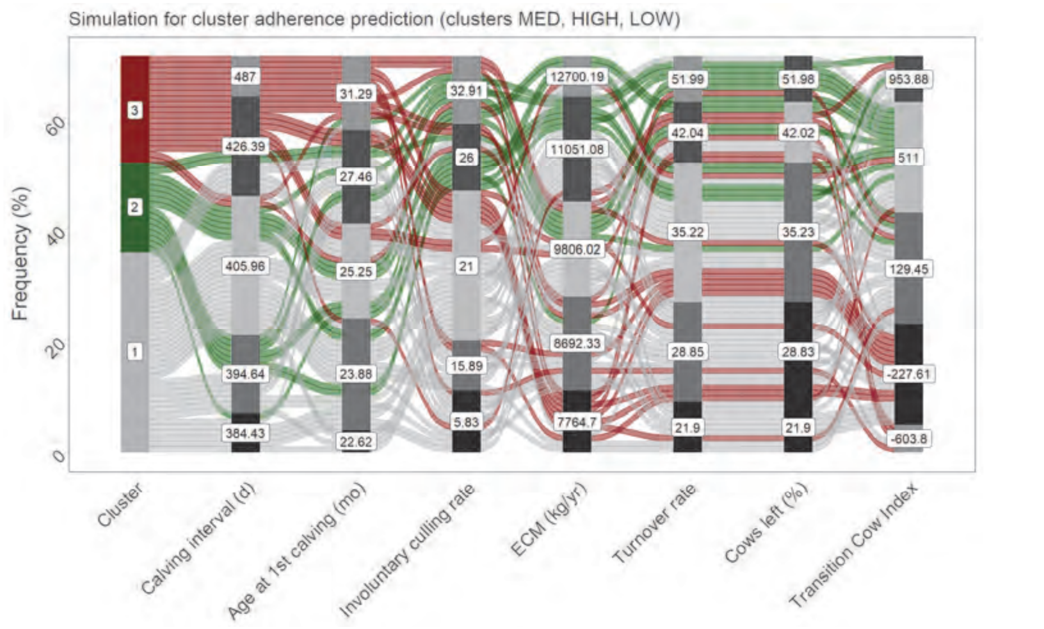

Figure 1. Model response, as shown by an alluvial diagram with predicted cluster adherence as affected by the most important predictors ranked by their variable importance score. Nodes on the left denote the Medium (1), High (2), and Low (3) performance cluster; nodes for predictors with respective means were fixed to five nodes based on the frequency distribution (y-axis).

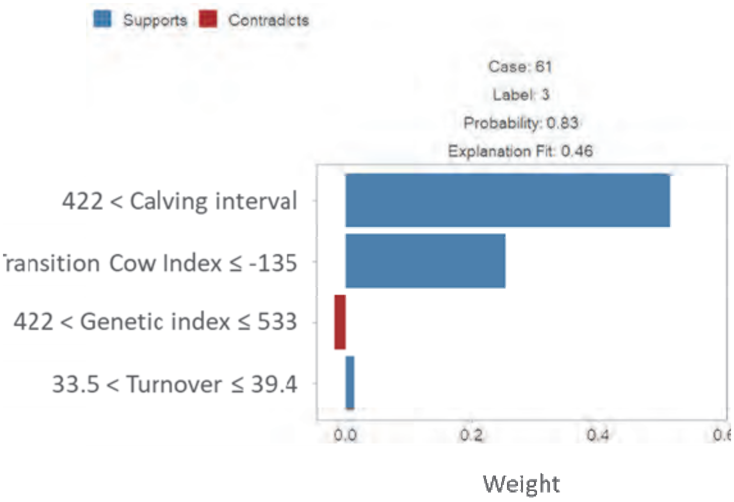

Figure 2. Contribution of variables of a randomly selected herd pertaining to the Low performing cluster (Cluster 3) based on a model agnostic approach.

making by highlighting possible similarities or discrepancies among herds within the same cluster and by pinpointing to potential deficiencies for a producer to focus on.

Conclusion

The findings presented here suggest that mining DHI data can give a valuable insight into best herd management practices and can be used to highlight opportunities for improvements. DHI records have a high potential for decision making support. Due to the type of records at hand, a more flexible approach using machine learning is recommended to deal with the large data volume and variety. Clustering allowed to identify 3 main herd clusters based on the overall herd performance using DHI data. High performing herds differentiated themselves from low performing herds mainly by adopting an optimal transition management, overall good reproduction performance and high attention given to the non-producing animals despite a potential for improvement in early lactation management. Decision tree induction further allowed to identify the most important DHI indicators explaining the adherence of a herd to one of the three clusters. This ultimately allowed to establish high-level recommendations and visualize which performance indicators a herd might want to focus on to improve overall herd performance. A local interpretable model fitted to single decision tree predictions allowed to add interpretability to the decision tree model and help with decision making by highlighting specific possible deficiencies within a herd.

Acknowledgement

This project was funded by Agriculture and Agri-Food Canada (Ottawa, ON, Canada) and Milk 2020 (Sussex, NB, Canada) through the Canadian Agricultural Partnership joint research program.

References

Abbott, D. 2014. Applied predictive analytics: Principles and techniques for the professional data analyst. Wiley and Sons, Indianapolis, IN.

Allaire, J.J., J. Horner, Y. Xie, V. Marti and N. Porte. 2019. markdown: Render Markdown with the C Library 'Sundown'. R package version 1.1. https://CRAN.R-project.org/package=markdown.

Breiman, L., J.H. Friedman, R.A. Olsen and C.J. Stone. 1984. Classification and regression trees. Wadsworth and Brooks, Monterey, CA.

Dominiak, K.N. and A.R. Kristensen. 2017. Prioritizing alarms from sensor-based detection models in livestock production - A review on model performance and alarm reducing methods. Computers and Electronics in Agriculture 133: 46–67. https://doi.org/10.1016/j.compag.2016.12.008.

Foss. 2009. Foss Application Note 35. MilkoScan FT+ Ketosis Calibrations. Foss, Hillerød, Denmark.

Heald, C.W., T. Kim, W.M. Sischo, J.B. Cooper and D.R. Wolfgang. 2000. A computerized mastitis decision aid using farm-based records: an artificial neural network approach. Journal of Dairy Science 83(4): 711–720. https://doi.org/10.3168/jds.S0022-0302(00)74933-2.

Kuhn, M. 2021. caret: Classification and Regression Training. R package version 6.0-88. https://CRAN.R-project.org/package=caret.

Kuznetsova, A., P. B. Brockhoff and R.H.B. Christensen. 2017. lmerTest Package: Tests in linear mixed effects models. Journal of Statistical Software 82(13): 1–26. doi: 10.18637/jss.v082.i13.

Lactanet, 2021. Portrait des fermes laitières québécoises en matière de logement et de gestion des animaux. Lactanet: https://lactanet.ca/wp-content/uploads/2022/01/brochure_portrait-ferme-qc_021v11.pdf (accessed on May 2022).

Le, S., J. Josse and F. Husson. 2018. FactoMineR: An R package for multivariate analysis. Journal of Statistical Software 25: 10.18637/jss.v025.i01.

Nordlund, K., 2006. Transition Cow Index. In: Proceedings of the 39th Annual Conference of American Association Bovine Practitioners, St. Paul, MN, USA, 20–24 September 2006; American Association of Bovine Practitioners, Ashland, OH, USA, p 139–143.

Premraj, R. 2015. mailR: A Utility to Send Emails from R. R package version 0.4.1. https://CRAN.R-project.org/package=mailR.

Ribeiro, M.T., S. Singh and C. Guestrin. 2016. Why should I trust you? Explaining the predictions of any classifier. arXiv:1602.04938 [cs.LG].

Rokach, L. and O. Maimon. 2005. Decision trees. In: Data Mining and Knowledge Discovery Handbook, Maimon, O. and L. Rokach (Editors), Springer, Boston, MA, p 165–192.

Tang, F. and H. Ishwaran. 2017. Random forest missing data algorithms. Statistical Analysis and Data Mining 10 : 363–377. https://doi.org/10.1002/sam.11348.

van Buuren, S. 2012. Flexible imputation of missing data. CRC Chapman and Hall, Boca Raton, FL.