10. Data Analytics What Can New Analyses Techniques Bring for Better Farm Results 2

Title presentation

Machine-learning based prediction of test day milk yield using historical data of the previous lactation.

Author(s)

M. Salamone, I. Adriaens ,G. Opsomer, H. Atashi, A. Aernouts & Hostens, M.

Institution for which the first author of this abstract is working

Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium.

Department of Biosystems, Division Animal and Human Health Engineering, KU Leuven, Campus Geel, 2440 Geel, Belgium.

Abstract

Routine data collection on dairy farms has mainly been used to predict the future within the ongoing lactation of an animal. We attempt to use current lactation data with a random forest regression model to predict the milk yield on the first test day of the subsequent lactation. The data was provided by the MmmooOgle™ platform. The dataset represented 54 082 lactation records coming of 32 530 multiparous animals from 102 herds. The features of the model were defined as the milk yield and the days in milk (DIM) of the eight first test days from the previous lactation. Additionally, the DIM of the first test day from the next lactation was used as a feature. The response variable of our model was the milk yield of the first test day of the next lactation (kgTD1X+1) The data manipulation and the development of the model were done on the High Performance Computer of Ghent University while running the Apache Spark™ 3.0.0 framework. The random forest regression algorithm used the function from the MLlib integrated package. A first layer of data selection was applied to remove outliers: DIM first test ≤ 60, test day records > 7, Age at first calving < 5 years, Calving interval < 450 days. With the second layer of data selction quality of the data was improved, by removing the perturbed lactations using the MilkBot® model. The filtered dataset was randomly split on animal level into an external validation set (20%), and a model set (80%). The latter dataset was used in a grid search to identify the optimal hyperparmeter and to train the final nextMILK model (depth = 25, trees = 125). The performance evaluation of the model was done on the external validation set. The final nextMILK yielded an RMSE of 5.79 an R^2 of 0.56 and an MAE of 4.36. It was compared with a set of benchmark models, where it proved to be more accurate then the benchmark.This study illustrates the methodology and performance of a machine learning model for early lactation production of dairy cattle using historical data.