MACHINE-LEARNING BASED PREDICTION OF TEST DAY MILK YIELD USING HISTORICAL DATA OF THE PREVIOUS LACTATION.

PhD Researcher: Matthieu Salamone
INTRODUCTION
THE INITIATION OF THE LACTATION
THE TRANSITION PERIOD

- Animal variation to response during challenge period
- Identification of at-risk animals: crucial
  - Animal welfare
  - Economic
  - Efficiency ~ ecological footprint
RESEARCH QUESTION

- Data driven identification $\rightarrow$ Prediction of unperturbed production $\leftrightarrow$ Realised production
- Capacity to predict production in the transition period
  - Current models (Wood, Wilmink,…) $\rightarrow$ Possibly influenced by transition diseases
  - Usage of historic test day data $\rightarrow$ develop & validate a new predictive model
THE INPUT DATA

- Data Source ➔
  - Test-Day (TD) records, calving dates, breeding dates and dry-off dates, Lactation Curve Parameters (MilkBot):
    - 102 herds
    - 32 530 Animals (Multiparous)
    - 54 082 Lactation pairs
DATA SPLICING

Dataset

- 80% - Prediction set
- 20% - Test set

- Splicing at Animal Level
  - Animal data either in Prediction or in Test Set
  - Reduces Possible bias in Performance evaluations
FEATURE SET

- Difference in collection ease and Quality standards ➔ Split feature set
- Working with 3 features set = Evaluation of added benefit more feature

### PRODUCTION

<table>
<thead>
<tr>
<th>DIM TD1_{x+1}</th>
<th>DIM TD1_x</th>
<th>DIM TD2_x</th>
<th>DIM TD3_x</th>
<th>DIM TD4_x</th>
<th>DIM TD5_x</th>
<th>DIM TD6_x</th>
<th>DIM TD7_x</th>
<th>DIM TD8_x</th>
<th>kg TD1_x</th>
<th>kg TD2_x</th>
<th>kg TD3_x</th>
<th>kg TD4_x</th>
<th>kg TD5_x</th>
<th>kg TD6_x</th>
<th>kg TD7_x</th>
<th>kg TD8_x</th>
<th>M21</th>
<th>M75</th>
<th>M305</th>
</tr>
</thead>
</table>

### REPRODUCTION

<table>
<thead>
<tr>
<th>Age At First Calving</th>
<th>Calving Interval</th>
<th>Days Dry</th>
<th>Days Open</th>
<th>Days In Milk</th>
<th>Δ Herd Age at First Calving</th>
<th>Δ Herd Calving Interval</th>
<th>Δ Herd Days Dry</th>
<th>Δ Herd Days Open</th>
<th>Δ Herd Days In Milk</th>
</tr>
</thead>
</table>

### HERD

<table>
<thead>
<tr>
<th>Δ Herd M21</th>
<th>Δ Herd M75</th>
<th>Δ Herd M305</th>
<th>Δ Herd Milk Minimum</th>
<th>Δ Herd Milk Maximum</th>
</tr>
</thead>
</table>
THE MACHINE LEARNING ALGORITHM

- Random Forest Regression (RF)
  - Creates set of decision trees
  - Parameters of the model:
    - Depth
    - # of trees
HYPERPARAMETER TUNING

– Grid search Hyperparameters=
  – Depth = 5, 10, 15, 20, 25
  – # Trees = 5, 25, 125, 250, 500
– 5 fold cross validation ➔ Optimal set = set with no significant difference with best performing set
FINAL MODELS NEXTMILK

Train final set of models → nextMILK models:

- The full training data set
- Optimal set of hyperparameters
- For each Feature Set
BENCHMARK MODELS

- Novel methodology
  - Evaluate our models
  - Performance of the models ↔ Complexity
- Investigate complexity
  - Simpler models ↔ Performance
  - Significant difference?
BENCHMARK I - HERD

\[ MILKP_{N+1} = MILKP_N \times (1 + \% \text{difference}_{\text{Herd}(N \rightarrow N+1)}) \]

- \( \% \text{difference}_{\text{Herd}} \) calculated over timeframes 0d - 75d and 0d - 305d
- Corrected for DIM
BENCHMARK II - ANIMAL

\[ MILKP_{N+1} = \overline{MILKP}_{N+1} \times (1 + \%\text{difference}_{\text{Animal} \rightarrow \text{Herd}(N)}) \]

- \( \%\text{difference}_{\text{Animal} \rightarrow \text{Herd}(N)} \) calculated over timeframes 0d - 75d and 0d - 305d
- Corrected for DIM
ANALYSIS OF BENCHMARK

– Predictions on test set of **nextMILK**
  – |Residuals| → ANOVA → post-hoc, one-sided paired t-test
– Performance metrics – RMSE, $R^2$, MAE
RESULTS & DISCUSSION
HYPERPARAMETER TUNING

Optimal Set:

- **FULL**: 125 trees, 25 deep
- **PRODUCTION HERD**: 250 trees, 25 deep
- **PRODUCTION**: 250 trees, 25 deep
**FINAL MODELS: NEXTMILK**

- Calculated on Test Set : 9 903 locations

Model performance indicator:
  - **Root Mean Squared Error**
  - **Mean Absolute Error**
  - **R²**

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>nextMILK - FULL</td>
<td>5.79</td>
<td>4.36</td>
<td>0.56</td>
</tr>
<tr>
<td>nextMILK – PRODUCTION HERD</td>
<td>6.01</td>
<td>4.53</td>
<td>0.52</td>
</tr>
<tr>
<td>nextMILK - PRODUCTION</td>
<td>6.03</td>
<td>4.52</td>
<td>0.52</td>
</tr>
</tbody>
</table>
BENCHMARK ANALYSIS

Statistical analysis:
- Significant difference between models
  ANOVA: P-value < 0.001
- All benchmark models have significantly higher residuals
- No Significant difference between nextMILK models.
# PERFORMANCE METRICS

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark I – HERD - 75</td>
<td>7.09</td>
<td>4.55</td>
<td>0.33</td>
</tr>
<tr>
<td>Benchmark I – HERD - 305</td>
<td>7.51</td>
<td>4.84</td>
<td>0.16</td>
</tr>
<tr>
<td>Benchmark II – ANIMAL -75</td>
<td>7.00</td>
<td>4.53</td>
<td>0.35</td>
</tr>
<tr>
<td>Benchmark II – ANIMAL -305</td>
<td>7.09</td>
<td>4.54</td>
<td>0.34</td>
</tr>
<tr>
<td>nextMILK - FULL</td>
<td>5.79</td>
<td>4.36</td>
<td>0.56</td>
</tr>
<tr>
<td>nextMILK – PRODUCTION HERD</td>
<td>6.01</td>
<td>4.53</td>
<td>0.52</td>
</tr>
<tr>
<td>nextMILK - PRODUCTION</td>
<td>6.03</td>
<td>4.52</td>
<td>0.52</td>
</tr>
</tbody>
</table>
CONCLUSION

- nextMILK models → show potential in predicting production in next lactation
- Further validation necessary → Data driven identification tool
Matthieu Salamone
PhD Student

DEPARTMENT OF REPRODUCTION, OBSTETRICS AND HERD HEALTH

E matthieu.salamone@ugent.be
T +32 9 264 75 62

www.ugent.be