Optimizing Disease Risk Classification: A Data-Integrative Approach

Caspar Matzhold[1], Katharina Schodl[1], Christa Egger-Danner[1]

ZuchtData EDV-Dienstleistungen GmbH, Dresdner Str. 89, 1200 Vienna, Austria

Background

Goal: Optimize models to classify cows at risk for specific diseases through **data integration**

Why: Models that enable targeted interventions can reduce health risks and support farm management practices, leading to better outcomes for specific farm systems

Key Considerations:

How can available data sources (dynamic, static) be integrated to improve disease risk prediction accuracy?

Study Background

Health Event of interest: **Ketosis**

Main Data Source: DHI (Dairy Herd Improvement)

- Fat-to-protein ratio, milk yield, … provides valuable information for predicting metabolic disorders
- **KetoMIR**, a mid-infrared-based technique, shows promising results for predicting ketosis

Additional Data Sources:

- Genetics, feeding, temperature, …

Research Question

1. **DHI**
   How accurately can we predict ketosis outcomes using **DHI & KetoMIR** data?

2. **DHI + Additional Data Sources**
   Does the inclusion of cow- and farm-level information further improve the model's accuracy?

Ketosis

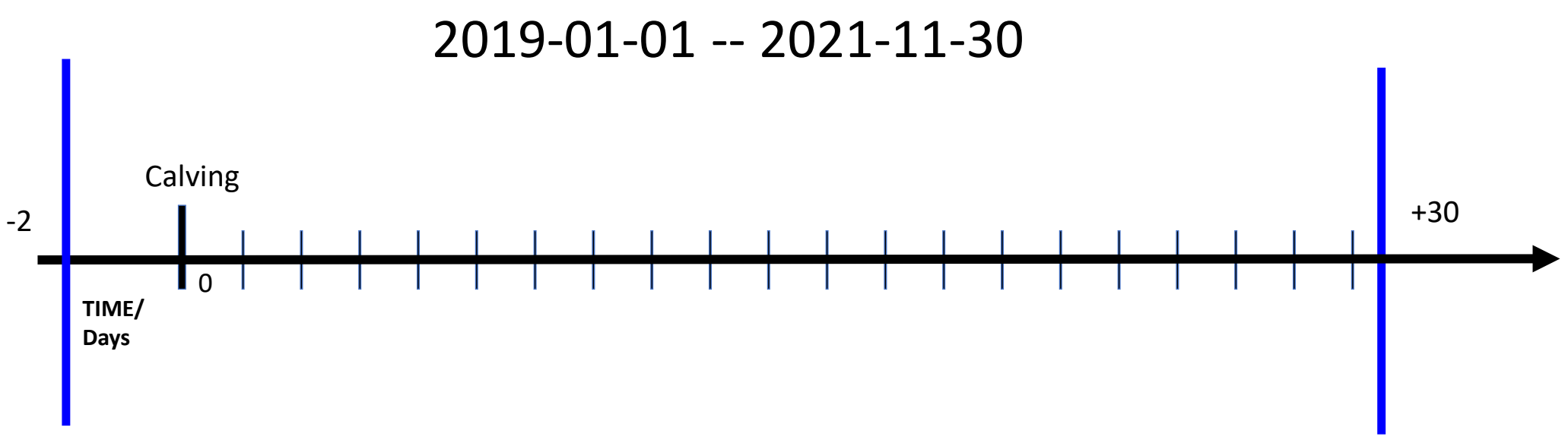
Data Sources for Health Classification:

To classify a positive event (indicating sickness), we derived health information from three key data sources:

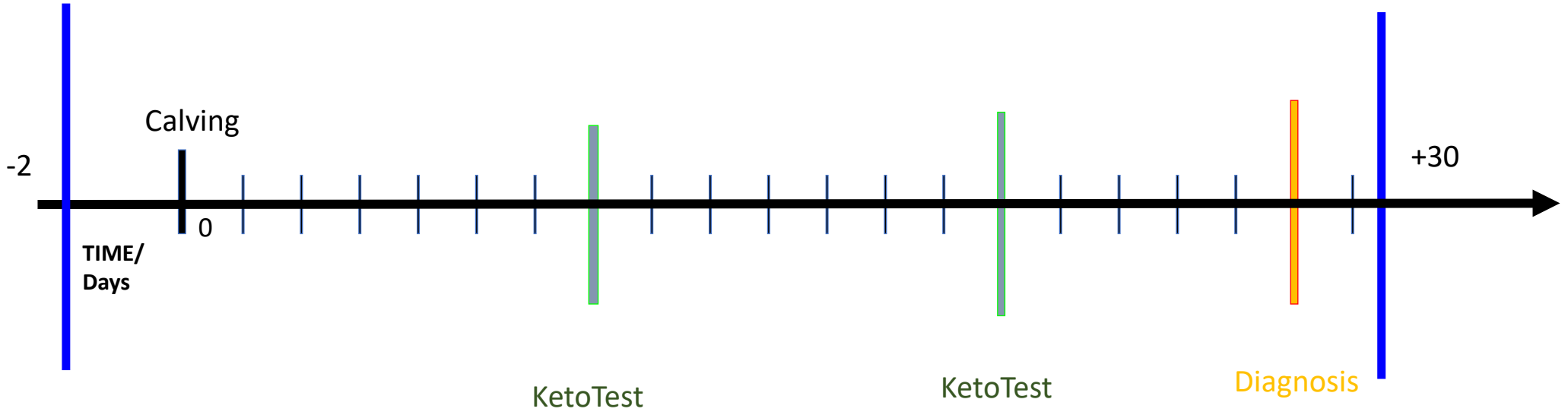
- **Veterinarian Diagnosis**
- **Cullings due to Metabolic Disorders**
- **KetoTest**
  - Blood & Milk

Test Type	Code	Value Range
KTB (Keto Blood Test)	01	0 - 1.2
	02	1.3 - 2.9
	03	≥ 3
KEB (Keto Blood Test)	01 - 03	0 - 1.2
	04 - 08	1.3 - 2.9
	09	≥ 3
KTM (Keto Milk Test)	-	Same as KTB/KEB
KetoMIR	Class 1	0 to 0.5
	Class 2	>0.5 to 0.75
	Class 3	>0.75

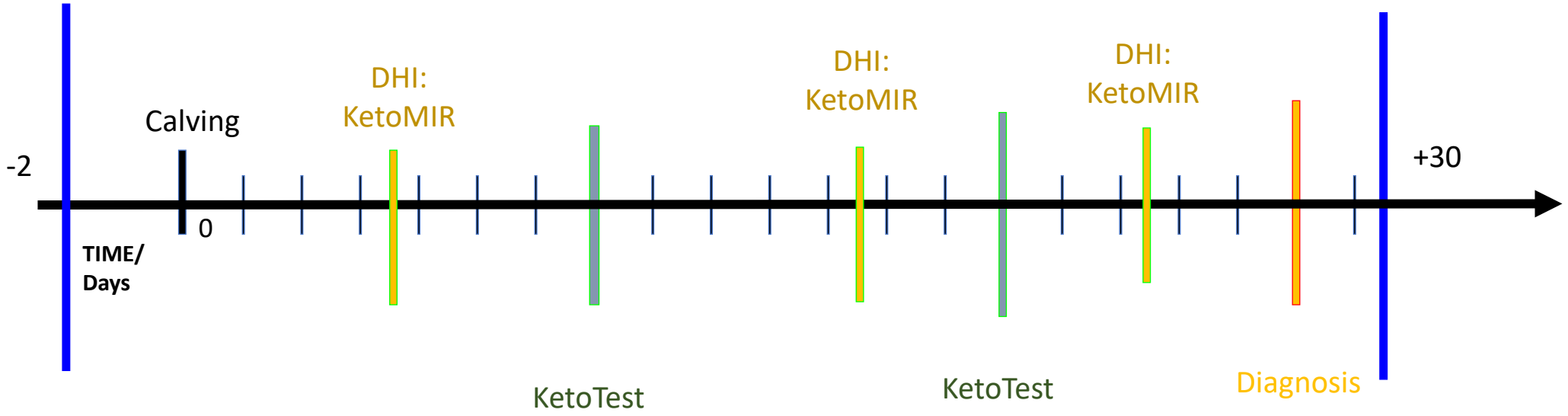
Health Outcome: Timeframe Consideration



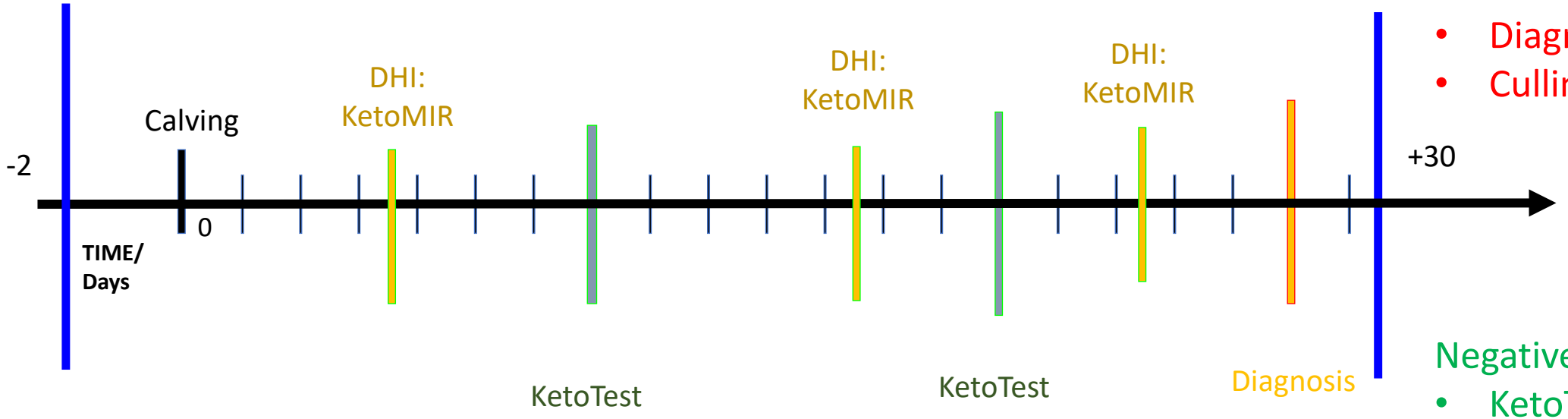
Health Outcome: Timeframe Consideration



Health Outcome: Timeframe Consideration



Health Outcome: Timeframe Consideration



Positive Event:

- KetoTest Positive OR
- Diagnosis OR
- Culling

Negative Event:

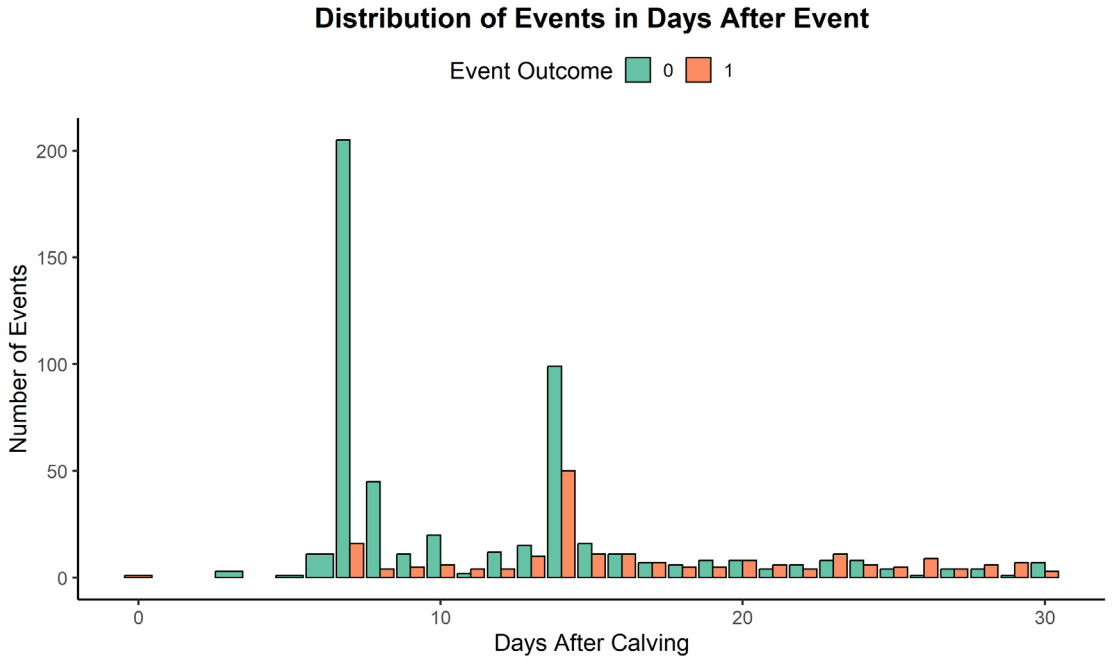
- KetoTest Negative
- No Diagnosis
- No Culling

Descriptives

Data Source	Observations	Farms	Cows	Positive Events	Negative Events	Positive KetoMIR Events	Negative KetoMIR Events
DHI	735	150	689	208	527	137	598

AV Time between Calving and

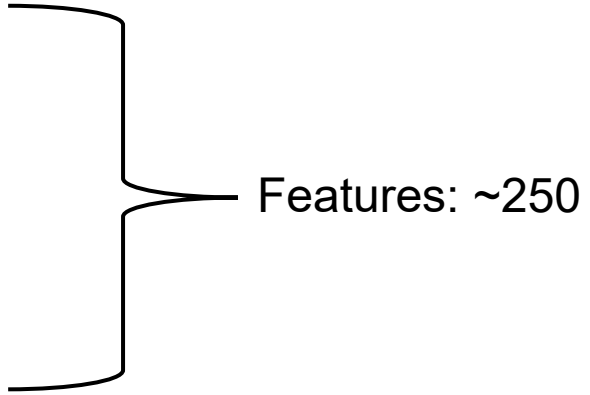
- DHI: 10.13 days
- Event: 13.02 days



Data: Manipulation & Aggregation

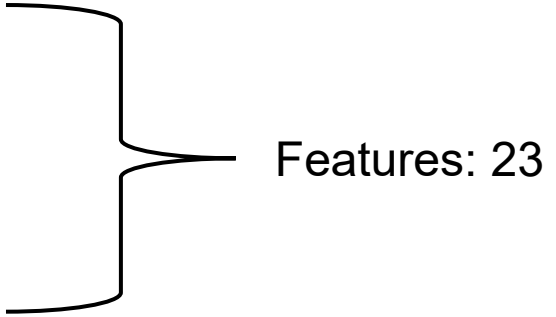
Data Sources:

- Dairy Herd Improvement (DHI) assessments
- Farm Survey
- National Weather Service (ZAMG)
- National Disease Registry
- National Cattle Database (RDV)
- National Cow Registry



Feature Reduction Techniques:

- Knowledge-based feature selection
- Correlation analysis (Multicollinearity)
- Information completeness
- Heterogeneity in information



Reduced Dataset

Data Source	Features		Number of Features
Basic Information	id.cow, id.farm, event, date		4
DHI	KetoMIR, Milk Yield, Fat-Protein-Ratio, Number of Lactations		4
Genetics	Breeding Value		1
Health	Health History, Ketosis Prophylaxis		2
Weather	Low-temperature-days-count, High-temperature-days-count		2
Farm	Management	Management: Organic, Overall Cow Count, Milk Yield Average	14
	Technological System	Sensor System	
	Housing System	Ventilation System, Cooling System	
	Feeding Management	Feeding Robot, Forage Type: Mixed Ratio, Forage Type: Template, Forage Type: Mixing Wagon Proportion of Concentrates between 35% - 45%, Proportion of Concentrates greater than 45%, Roughage: min. 3cm length, Feeding Group: Uniform	

Methods

Models

Logistic Regression:

- Assumes linear relationships
- Simple, interpretable

Random Forest (RF):

- Ensemble method, handles non-linear data
- Aggregates multiple decision trees

XGBoost:

- Boosting ensemble, sequentially corrects errors
- Handles complex non-linear data

Model Metrics

Sensitivity (True Positive Rate):

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Specificity (True Negative Rate):

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

F1-Score (Harmonic Mean of Precision and Sensitivity):

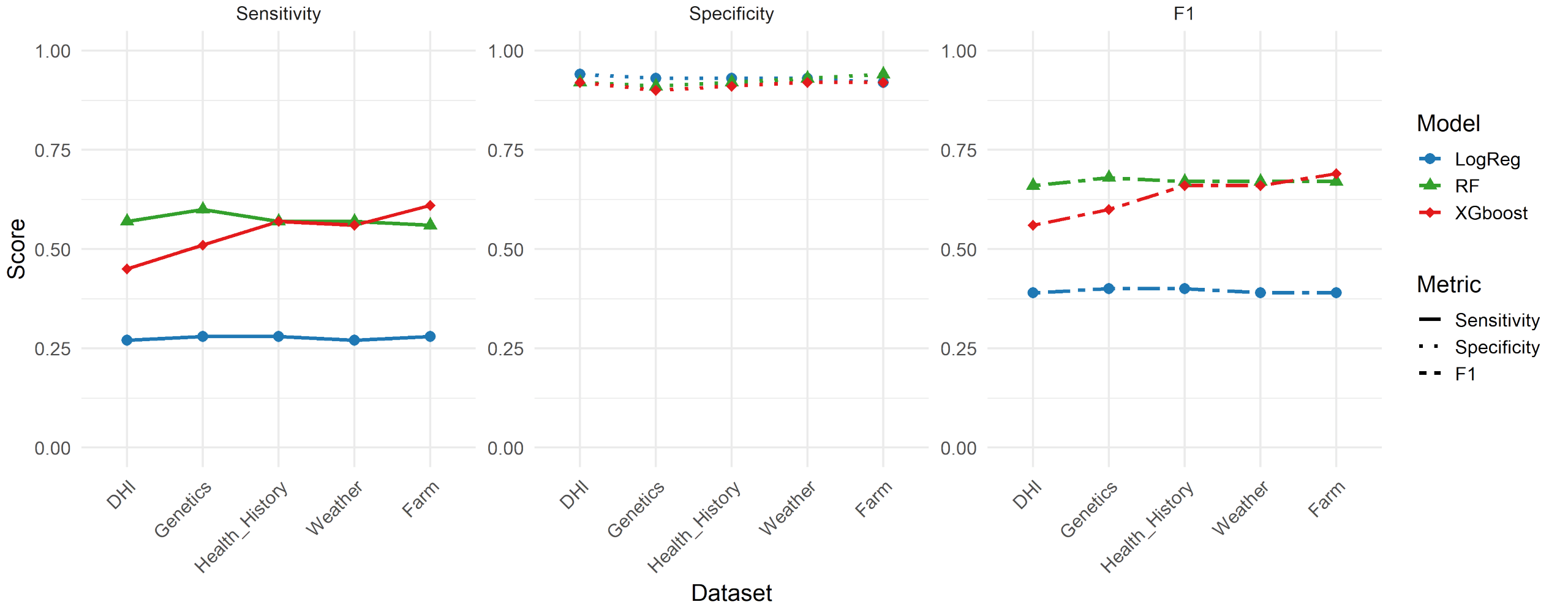
$$\text{F1-Score} = \frac{2 \times \text{Sensitivity} \times \text{Precision}}{\text{Sensitivity} + \text{Precision}}$$

Results

RINDERZUCHT AUSTRIA Innovation

ZUCHT DATA

Impact of Data Integration on Model Performance



Data Value: F1-Score

Data	Model	Sensitivity	Specificity	F1
DHI	LogReg	0.27	0.94	0.39
DHI	RF	0.57	0.92	0.66
DHI	XGBoost	0.45	0.92	0.56
Genetics	LogReg	0.28	0.93	0.4
Genetics	RF	0.6	0.91	0.68
Genetics	XGBoost	0.51	0.9	0.60
Health_History	LogReg	0.28	0.93	0.4
Health_History	RF	0.57	0.92	0.67
Health_History	XGBoost	0.57	0.91	0.66
Weather	LogReg	0.27	0.93	0.39
Weather	RF	0.57	0.93	0.67
Weather	XGBoost	0.56	0.92	0.66
Farm	LogReg	0.28	0.92	0.39
Farm	RF	0.56	0.94	0.67
Farm	XGBoost	0.61	0.92	0.69

Data Value: F1-Score

Data	Model	Sensitivity	Specificity	F1
DHI	LogReg	0.27	0.94	0.39
DHI	RF	0.57	0.92	0.66
DHI	XGBoost	0.45	0.92	0.56
Genetics	LogReg	0.28	0.93	0.4
Genetics	RF	0.6	0.91	0.68
Genetics	XGBoost	0.51	0.9	0.60
Health_History	LogReg	0.28	0.93	0.4
Health_History	RF	0.57	0.92	0.67
Health_History	XGBoost	0.57	0.91	0.66
Weather	LogReg	0.27	0.93	0.39
Weather	RF	0.57	0.93	0.67
Weather	XGBoost	0.56	0.92	0.66
Farm	LogReg	0.28	0.92	0.39
Farm	RF	0.56	0.94	0.67
Farm	XGBoost	0.61	0.92	0.69

Data Value: F1-Score

Data	Model	Sensitivity	Specificity	F1
DHI	LogReg	0.27	0.94	0.39
DHI	RF	0.57	0.92	0.66
DHI	XGBoost	0.45	0.92	0.56
Genetics	LogReg	0.28	0.93	0.4
Genetics	RF	0.6	0.91	0.68
Genetics	XGBoost	0.51	0.9	0.60
Health_History	LogReg	0.28	0.93	0.4
Health_History	RF	0.57	0.92	0.67
Health_History	XGBoost	0.57	0.91	0.66
Weather	LogReg	0.27	0.93	0.39
Weather	RF	0.57	0.93	0.67
Weather	XGBoost	0.56	0.92	0.66
Farm	LogReg	0.28	0.92	0.39
Farm	RF	0.56	0.94	0.67
Farm	XGBoost	0.61	0.92	0.69

Data Value: F1-Score

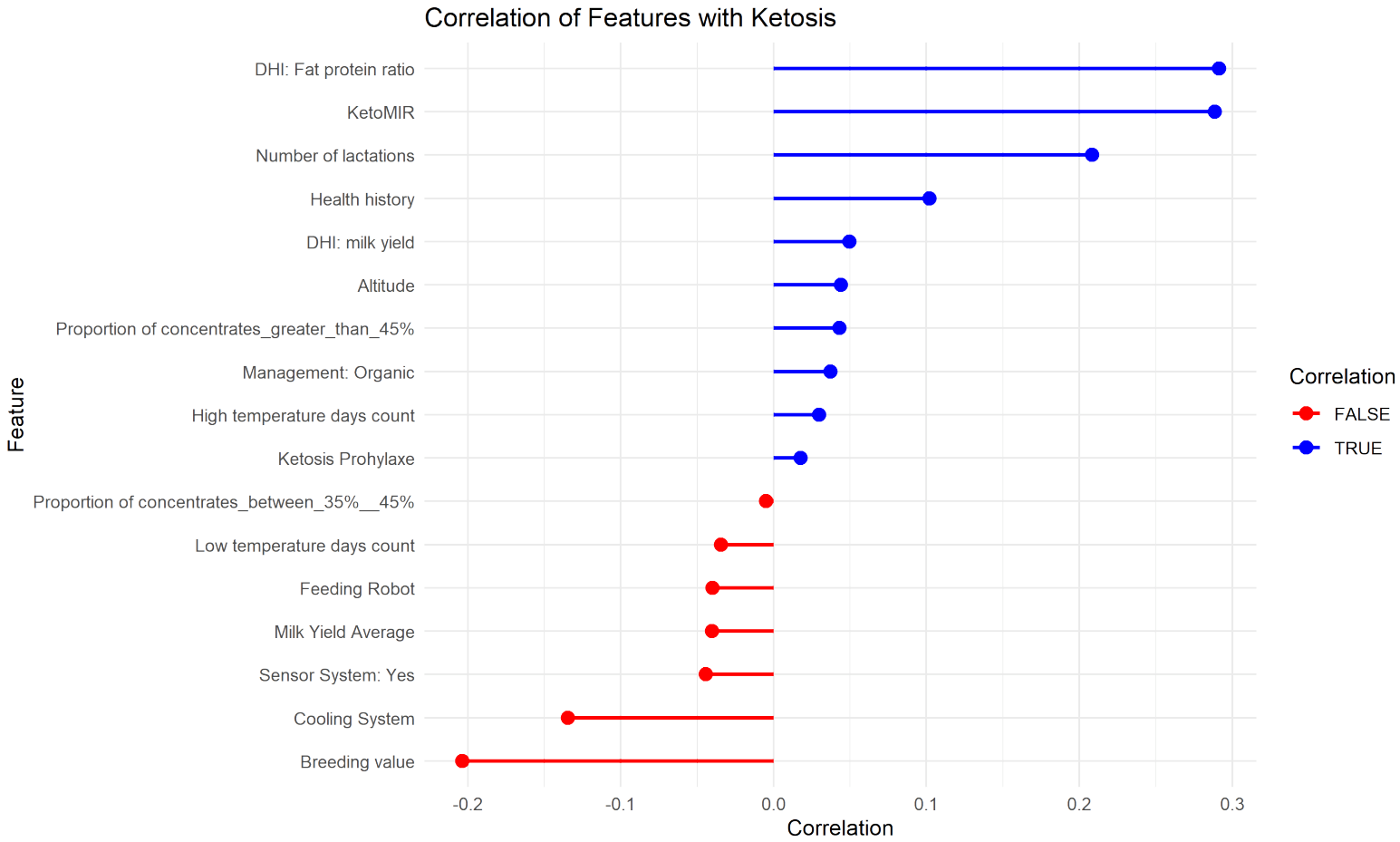
Data	Model	Sensitivity	Specificity	F1
DHI	LogReg	0.27	0.94	0.39
DHI	RF	0.57	0.92	0.66
DHI	XGBoost	0.45	0.92	0.56
Genetics	LogReg	0.28	0.93	0.4
Genetics	RF	0.6	0.91	0.68
Genetics	XGBoost	0.51	0.9	0.60
Health_History	LogReg	0.28	0.93	0.4
Health_History	RF	0.57	0.92	0.67
Health_History	XGBoost	0.57	0.91	0.66
Weather	LogReg	0.27	0.93	0.39
Weather	RF	0.57	0.93	0.67
Weather	XGBoost	0.56	0.92	0.66
Farm	LogReg	0.28	0.92	0.39
Farm	RF	0.56	0.94	0.67
Farm	XGBoost	0.61	0.92	0.69

Data Value: F1-Score

Data	Model	Sensitivity	Specificity	F1
DHI	LogReg	0.27	0.94	0.39
DHI	RF	0.57	0.92	0.66
DHI	XGBoost	0.45	0.92	0.56
Genetics	LogReg	0.28	0.93	0.4
Genetics	RF	0.6	0.91	0.68
Genetics	XGBoost	0.51	0.9	0.60
Health_History	LogReg	0.28	0.93	0.4
Health_History	RF	0.57	0.92	0.67
Health_History	XGBoost	0.57	0.91	0.66
Weather	LogReg	0.27	0.93	0.39
Weather	RF	0.57	0.93	0.67
Weather	XGBoost	0.56	0.92	0.66
Farm	LogReg	0.28	0.92	0.39
Farm	RF	0.56	0.94	0.67
Farm	XGBoost	0.61	0.92	0.69

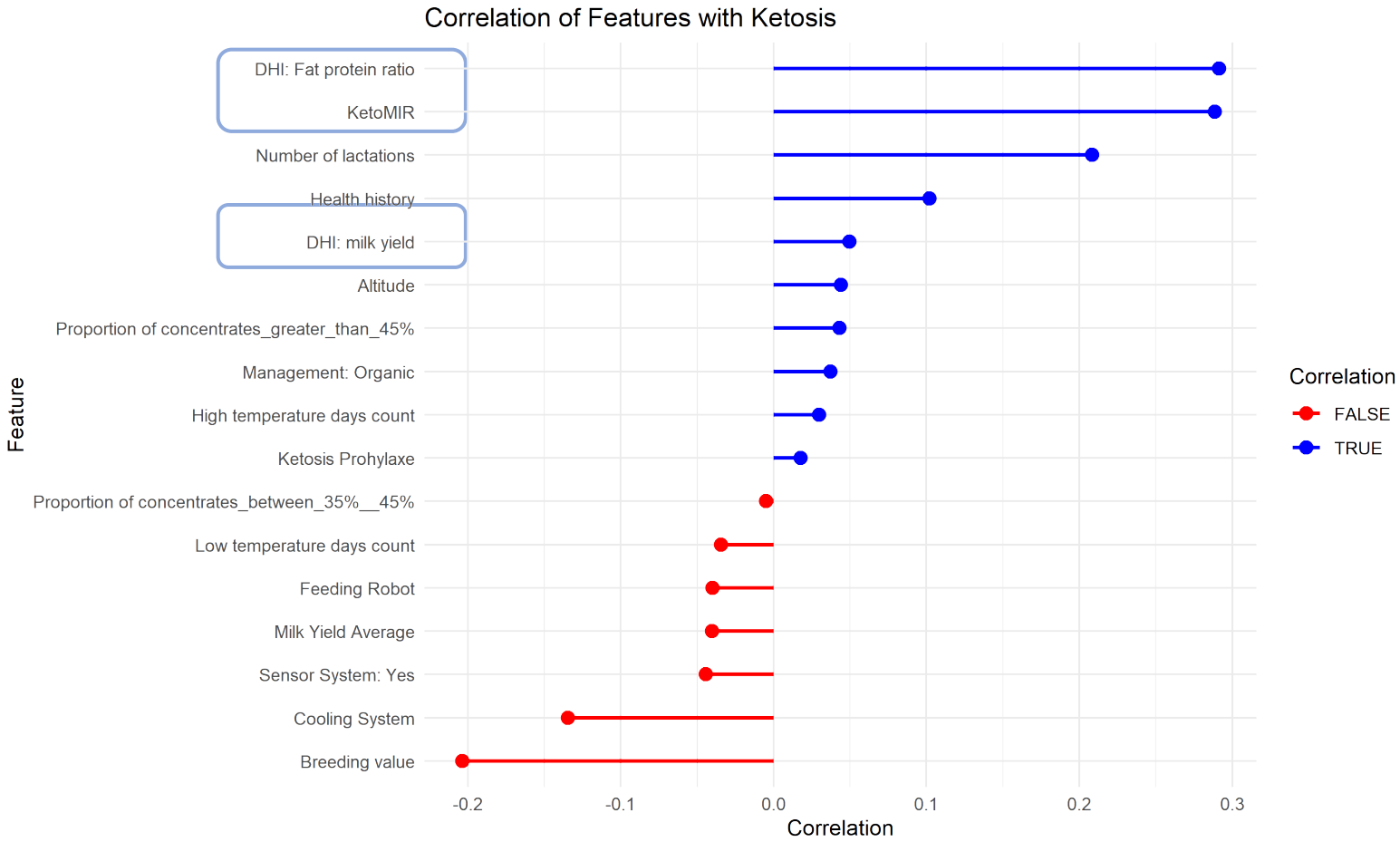
Feature Importance Rank and Correlations

Feature	LogReg	RF	XGBoost
DHI: Fat-Protein-Ratio	1	1	1
DHI: Milk Yield	11	2	2
Breeding Value	9	3	3
Milk Yield Average		4	4
Overall Cow Count		6	5
High-temperature-days-count		7	6
Low-temperature-days-count		8	7
Number of Lactations	2	5	8
Altitude	8	9	9
KetoMIR	5	10	10
Cooling System	4	12	11
Sensor System: Yes	3	11	12
Roughage: min. 3cm length	7	13	13
Ketosis Prophylaxis	12	14	14
Feeding Robot			15
Management: Organic	10		
Feeding Group: Uniform	6		
Proportion of Concentrates between 35% - 45%	14		
Proportion of Concentrates greater than 45%	13		
Ventilation System	15	15	



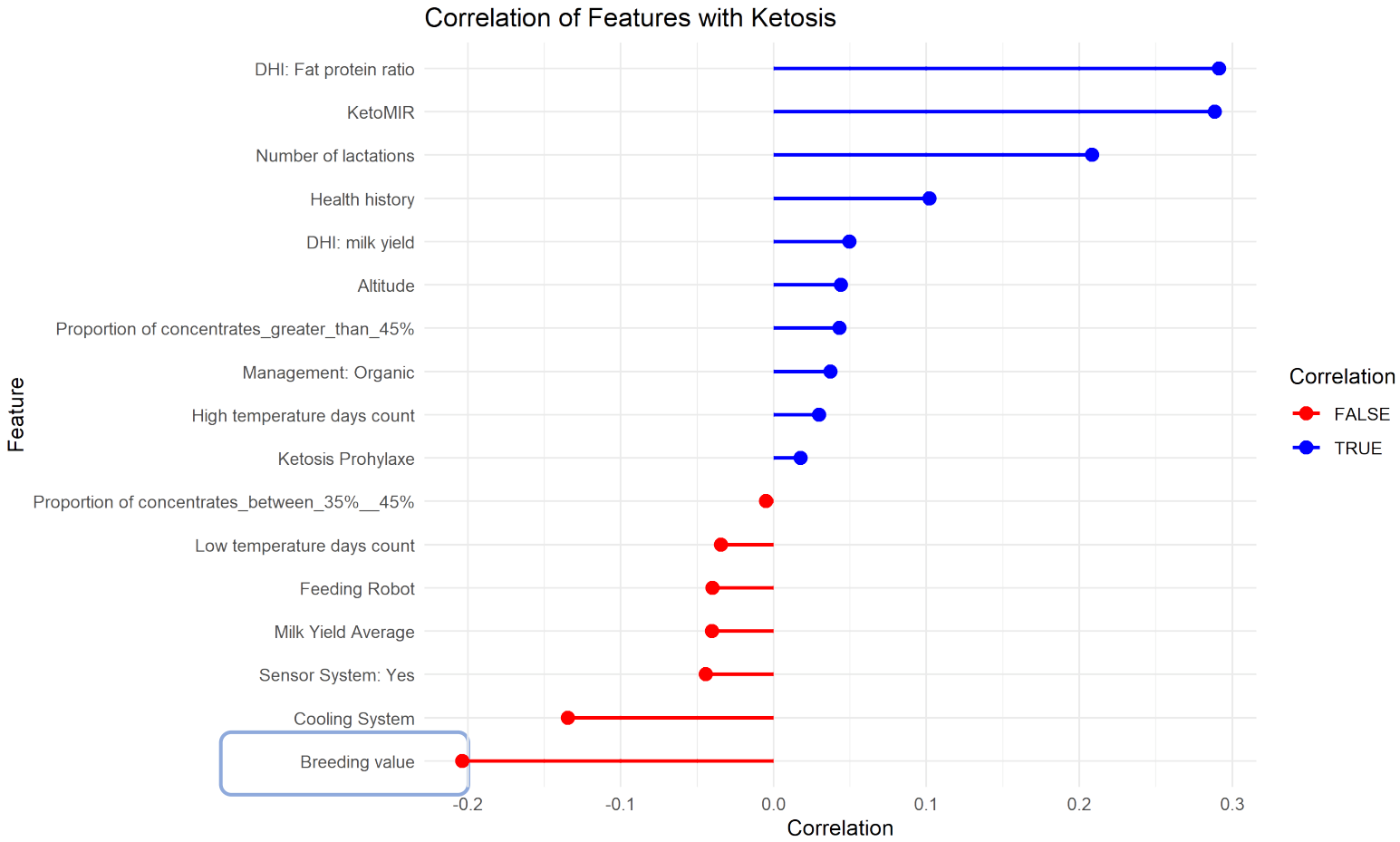
Feature Importance Rank and Correlations

Feature	LogReg	RF	XGBoost
DHI: Fat-Protein-Ratio	1	1	1
DHI: Milk Yield	11	2	2
Breeding Value	9	3	3
Milk Yield Average		4	4
Overall Cow Count		6	5
High-temperature-days-count		7	6
Low-temperature-days-count		8	7
Number of Lactations	2	5	8
Altitude	8	9	9
KetoMIR	5	10	10
Cooling System	4	12	11
Sensor System: Yes	3	11	12
Roughage: min. 3cm length	7	13	13
Ketosis Prophylaxis	12	14	14
Feeding Robot			15
Management: Organic	10		
Feeding Group: Uniform	6		
Proportion of Concentrates between 35% - 45%	14		
Proportion of Concentrates greater than 45%	13		
Ventilation System	15	15	



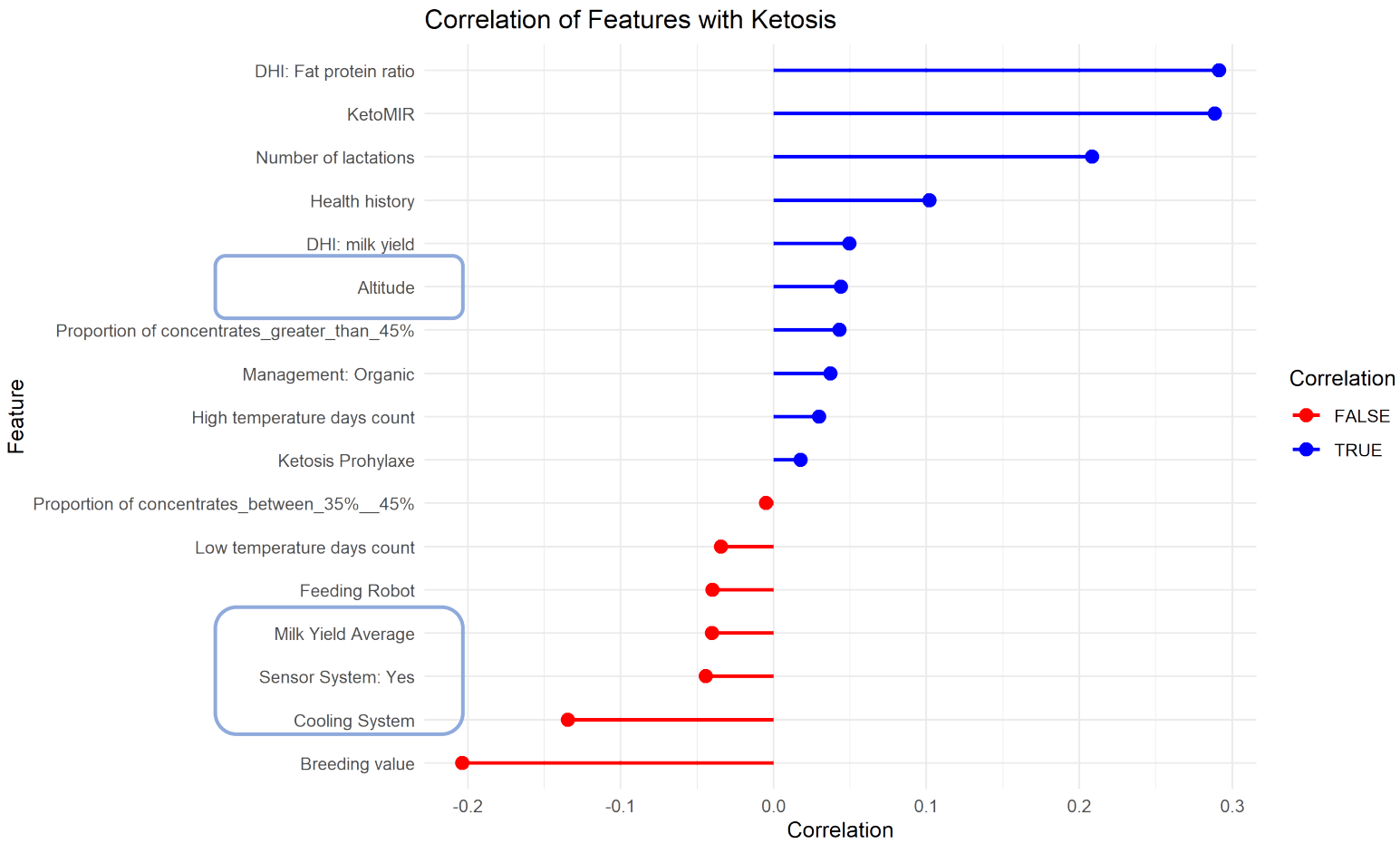
Feature Importance Rank and Correlations

Feature	LogReg	RF	XGBoost
DHI: Fat-Protein-Ratio	1	1	1
DHI: Milk Yield	11	2	2
Breeding Value	9	3	3
Milk Yield Average		4	4
Overall Cow Count		6	5
High-temperature-days-count		7	6
Low-temperature-days-count		8	7
Number of Lactations	2	5	8
Altitude	8	9	9
KetoMIR	5	10	10
Cooling System	4	12	11
Sensor System: Yes	3	11	12
Roughage: min. 3cm length	7	13	13
Ketosis Prophylaxis	12	14	14
Feeding Robot			15
Management: Organic	10		
Feeding Group: Uniform	6		
Proportion of Concentrates between 35% - 45%	14		
Proportion of Concentrates greater than 45%	13		
Ventilation System	15	15	



Feature Importance Rank and Correlations

Feature	LogReg	RF	XGBoost
DHI: Fat-Protein-Ratio	1	1	1
DHI: Milk Yield	11	2	2
Breeding Value	9	3	3
Milk Yield Average		4	4
Overall Cow Count		6	5
High-temperature-days-count		7	6
Low-temperature-days-count		8	7
Number of Lactations	2	5	8
Altitude	8	9	9
KetoMIR	5	10	10
Cooling System	4	12	11
Sensor System: Yes	3	11	12
Roughage: min. 3cm length	7	13	13
Ketosis Prophylaxis	12	14	14
Feeding Robot			15
Management: Organic	10		
Feeding Group: Uniform	6		
Proportion of Concentrates between 35% - 45%	14		
Proportion of Concentrates greater than 45%	13		
Ventilation System	15	15	



Conclusion

How accurately can we predict outcomes using only DHI data?

- DHI information achieves a prediction F1-Score ranging from 0.39 to 0.56, depending on the model
- DHI data is highly valuable for prediction, as shown by feature importance and correlations

Does the inclusion of cow- and farm-level information further improve the model's accuracy?

- Yes, data integration enhances the performance of ketosis classification models
- Using the full aggregated dataset with the best model, XGBoost, improves the F1-Score from 0.56 to 0.69 and Sensitivity from 0.45 to 0.61

Model Performance Insights:

- **Regression Models:** Logistic Regression shows limited ability to manage complex, multi-source data
- **XGBoost & Random Forest:** Both models perform similarly, with XGBoost showing the best results

- Attachment

Filtering – Sample Size

Filter	Negative	Positive Events	Positive Event Types
Base	10.253	2.950	Culling: 64, DIAG: 1,706, KEB: 7,598, KTB: 1,072, KTM: 2,711, TAB: 52
Year	9.226	1.764	Culling: 63, DIAG: 672, KEB: 7,396, KTB: 732, KTM: 2,112, TAB: 15
Lactation: 300 days	5.327	1.481	Culling: 50, DIAG: 610, Healthy: 5,327, KEB: 685, KTB: 70, KTM: 52, TAB: 14
Calving: -2/30 days	5.161	1.224	Culling: 27, DIAG: 418, Healthy: 5,161, KEB: 651, KTB: 66, KTM: 50, TAB: 12
KetoMIR before Event -22/1	**527**	**208**	**DIAG: 76, Healthy: 527, KEB: 99, KTB: 20, KTM: 11, TAB: 2**