BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Data to decisions

B. J. White¹, D. E. Amrine, and R. L. Larson

Beef Cattle Institute, Kansas State University, Manhattan, KS

ABSTRACT: Big data are frequently used in many facets of business and agronomy to enhance knowledge needed to improve operational decisions. Livestock operations collect data of sufficient quantity to perform predictive analytics. Predictive analytics can be defined as a methodology and suite of data evaluation techniques to generate a prediction for specific target outcomes. The objective of this manuscript is to describe the process of using big data and the predictive analytic framework to create tools to drive decisions in livestock production, health, and welfare. The predictive analytic process involves selecting a target variable, managing the data, partitioning

the data, then creating algorithms, refining algorithms, and finally comparing accuracy of the created classifiers. The partitioning of the datasets allows model building and refining to occur prior to testing the predictive accuracy of the model with naive data to evaluate overall accuracy. Many different classification algorithms are available for predictive use and testing multiple algorithms can lead to optimal results. Application of a systematic process for predictive analytics using data that is currently collected or that could be collected on livestock operations will facilitate precision animal management through enhanced livestock operational decisions.

Key words: big data, cattle, precision livestock

© The Author(s) 2018. Published by Oxford University Press on behalf of American Society of Animal Science. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

J. Anim. Sci. 2018.96:1531–1539
doi: 10.1093/jas/skx065

INTRODUCTION

Investigation of livestock production involves studies at many hierarchical levels of organization within biology, from atoms to the entire biosphere. The goal of collecting data in biologic systems is to generate more accurate descriptions of system aspects to inform decisions. Statistics are often used to evaluate data collected from complex biologic systems with substantial variability to identify trends, patterns, or interactions, which can lead to new knowledge. This

information can also improve the ability to make important decisions with greater positive and/or more predictable impacts. As computing power and data storage abilities have exponentially increased, so has the amount of data collected from commercial livestock production systems and experimental studies.

Although the ability to collect, process, and store large volumes of data has drastically increased, application of predictive analytic methodology in livestock environments is uncommon. Statistical analysis is often used to evaluate a relatively small sample size and determine the likelihood of finding a difference as large as observed when no difference actually exists. With advancements in connected sensors, the ability to monitor individual animals and their environments has provided the opportunity to collect large amounts of spatiotemporal and animal behavioral data.

Based on a presentation at the Big Data Analytics and Precision Animal Agriculture Symposium entitled “Data to decisions from remote cattle monitoring” held at the 2017 ASAS-CSAS Annual Meeting, July 12, 2017, Baltimore, MD.

¹Corresponding author: bwhite@vet.k-state.edu

Received September 14, 2017.

Accepted March 19, 2018

These large datasets afford the opportunity to evaluate beyond associations, and data can be used to classify observations into relevant subgroups or predict outcomes. The objective of this paper is to describe the process of using big data combined with predictive analytic methods to drive decisions in livestock production, health, and welfare. Specifically, this manuscript focuses on using operational data to create a framework for prediction and classification to provide new information to animal managers.

BIG DATA IN CATTLE PRODUCTION AND RESEARCH

The availability of inexpensive computing power, large capacity storage media, and internet connectivity have exponentially increased the amount of data collected from individuals and cohorts of livestock. Many cattle feeding operations have group and individual animal data collected from the time an animal is purchased at an auction market through carcass characteristic classifications collected at the abattoir. Combining individual animal records such as weights, treatments, and carcass characteristics with cohort-level information such as daily feed deliveries, diet ingredients, group weights, and movements provides significant amounts of data useful for analysis when the correct analytical framework is utilized (Theurer et al., 2015). The goal of any analytic project is to gain insight into the factors and interactions influencing a complex system and often to have impact on critical decisions (Abbott, 2014). Production datasets incorporating cohort and individual animal data can easily become large representing millions of observations. The advent of big data drives the need for decision tools to effectively transform the large volume of raw numbers into effective decision matrices.

In addition to typical production system data, multiple technologies can be utilized to remotely assess changes in animal wellness status (Theurer et al., 2013). These technologies can be used in a research setting or some technologies have been applied in commercial environments. Activity monitoring systems such as accelerometers or real-time location systems capture recursive numeric positions of individual animals in relatively short time-steps. These data can be used by algorithms to determine if these movements are within or outside expected ranges and to help identify patterns that are associated with physiologic changes such as estrus or disease. Cohort and individual animal data combined with individual animal

activity monitoring systems can provide big data pools valuable for both descriptive and predictive analytics.

STATISTICS AND PREDICTIVE ANALYTICS

Traditional inferential statistical methods are often used to evaluate an outcome of interest, to test potential associations with other variables, and in some cases, to provide predictions for new data (White et al., 2016). Properly designed controlled experiments use techniques to avoid bias and the collected data are frequently analyzed using traditional inferential statistical methods. Data must meet certain assumptions for many traditional statistical methods to be appropriately applied to study data. Inferential statistics are a valuable tool to evaluate potential associations among variables in well-designed experimental research trials. The assessment goal is often to determine the likelihood that observed differences among variables are due to random chance if no difference between these variables exists. This differs from the main objective of predictive analytics, which focuses on utilizing collected information to make a specific decision in the future.

Predictive analytic techniques are rarely listed as among common types of scientific study design or evaluation methods; however, the procedures framing predictive analytics follows the methodology for using scientific information to make decisions: develop a hypothesis/prediction, test the prediction, evaluate the results, interpret results, revise and repeat the process (Larson and White, 2015). Predictive analytics include several different analytic techniques including traditional statistics, machine learning, and data mining to discover meaningful patterns (Abbott, 2014). The conceptual distinction between predictive analytics and traditional inferential statistics is that the objective of predictive analytics is not to describe relationships among data, but rather to utilize data to create a model to be used on future data to predict an outcome of interest. The tools used for predictive analytics are most effectively applied to large datasets and these evaluations may not require the same assumptions regarding homogeneous variances and other traditional statistical assumptions. Similar to nonparametric statistics, many classification algorithms do not have an assumption regarding the underlying distribution of the covariates or prediction. The dataset must be representative of the population where the decision tool will be applied, but the value of the predictive method is based on direct assessment of accuracy of the prediction. The output of a predictive analytic tool is directly

assessed based on the accuracy of predictions using data with known outcomes and naive to the model (data not used in generating the tool).

Predictive analytics are focused on providing a prediction, and for some production and research questions, these models can have better performance compared to traditional statistical models. One example in livestock production data used a dataset of 108,931 daily milk yields and found that an artificial neural network, a type of predictive regression model, predicted daily milk yields more accurately than traditional regression models (Grzesiak et al., 2006). Another paper evaluated the ability to predict outcomes of diseased feedlot cattle and employed several classification algorithms including naive Bayesian classification, decision trees, random forests, and logistic regression. In this work, the authors found that logistic regression was rarely the optimal model when evaluating overall accuracy (Amrine et al., 2014). Classification algorithms differ in methodology or techniques used to minimize variation in outcomes of interest, and an ever-expanding library of potential machine learning algorithms are available for testing. Some models are better for different outcomes or datatypes, and unlike statistical modeling where a single model is selected prior to analysis, multiple models may be tested and the best algorithm determined based on the level of classification accuracy.

The predictive analytic process involves selecting a target variable, managing the data, partitioning the data, then creating algorithms, refining algorithms, and finally comparing accuracy of the created classifiers (Figure 1). Each step in the process is important to ensure outcomes have internal validity and provide the information desired to enhance subsequent decision making. Numerous

permutations exist related to specific models selected, data types included, and management of the raw dataset; however, the predictive analytic framework provided in Figure 1 provides a guide to facilitate systematic creation and evaluation of big data analysis.

DEFINING THE TARGET VARIABLE AND PREDICTIVE OUTCOME

One of the first steps in the predictive analytic process is identifying the target variable, which is the outcome of the data to be estimated or predicted (Abbott, 2014). The target variable should be carefully selected to provide information that will drive overall business decisions important to the financial sustainability of the operation. The type of data describing the target variable determines the structure of the predictive analytic problem and the appropriate model to be deployed. A classification problem deals with predicting a target variable that is qualitative and can take on one of two classifications (dead vs. alive, yes vs. no) or can be one of several ordinal classifications like a clinical illness score (CIS) or categories of yield grade (1, 2, 3, 4, 5). Regression problems deal with quantitative continuous measurements like BW or ADG. The data type of the target variable and variables used as predictors dictates the selection of the appropriate tool for prediction.

Once the target variable has been determined, the next step should be clearly defining the question to be answered. Narrowing the scope of an identified problem to an explicit hypothesis or question will inform the specific data needed, delimit the potential models, and define the level of accuracy that makes the model beneficial for making decisions. No model is perfect; therefore, reasonable

Figure 1. Schematic diagram representing systematic framework for predictive analytic workflow including data management, data partitioning, model building, refining predictive models, and accuracy assessment.

and clear performance goals will help guide the remainder of the project (Abbott, 2014). A clear, well-defined decision point will allow focus for the model to predict a specific piece of information that can provide a leverage point for operational decisions.

INITIAL DATA MANAGEMENT

Basic exploratory data analysis should be used to determine the extent of missing data and to visualize trends. Simple summary statistics, scatter plots, box plots, histograms, and two-by-two tables can be very helpful to guide model selection and to determine if the available data are sufficient to answer the hypothesis or question. Correlations between continuous variables can be assessed to better understand relationships among variables in the dataset and make decisions on inclusion or exclusion of specific variables.

Once the target variable is determined and basic data summarization has been used to evaluate the data structure, preprocessing of the data are required. Each analyst will have his or her preferred methods of preprocessing. In general, the process involves evaluating potential outliers, determining the level of missing data, and if it is appropriate, using methods to evaluate collinearity among variables (especially important if using traditional methods like linear regression). Many machine learning algorithms will handle missing data satisfactorily and is one benefit of using these methods rather than traditional statistical methods when evaluating production data collected from livestock systems; however, large amounts of missing data can cause variables to become unreliable predictors and missing data should be minimized if possible. Evaluation of why the data are missing is a time-consuming process, but should be considered and the most appropriate method to deal with the missing data should be considered before beginning the modeling process (Abbott, 2014). Not all missing data are detrimental to the prediction process as sometimes there are reasons that data are missing and the occurrence of missing data from specific animals, cohorts, or time can provide important information to the models. When missing data are problematic, these data can be imputed by a variety of methods including calculating most likely values based on data frequency or average values based on observations with similar attributes. Imputation of missing data should be used with caution as this may provide information in an area where data are truly unknown and

data would have to be imputed or calculated for new cases to be classified if these variables are also unknown when the classification algorithms are applied to new datasets.

Multicollinearity among two or more continuous variables can lead to biased coefficient estimation and loss of power (Yoo et al., 2014). When using traditional methods like multiple linear regression, multicollinearity can cause significant problems and should be strictly evaluated prior to model building. However, when using predictive methods such as random forests, because each tree is a random subset of the predictors, collinear variables can be kept in the data set as they may be used in different bootstrapped trees and each collinear variable could provide useful information (Hayes et al., 2015).

Predictive analytic models may benefit from the addition of new variables derived to capture predictive characteristics relative to target outcomes. These variables could represent completely new information added to the dataset from an external source (e.g., temporal weather data added to operational health dataset) or could be a combination of existing data in a new manner (e.g., creating a cumulative proportion of illness related to daily number of animals available to be ill added to an operational health dataset). For example, one paper used several classification tools to predict heat stress events and showed that addition of weather information augmented model performance compared to only evaluating data directly collected on infrared evaluation of animal surface temperatures (Unruh et al., 2017). Generating new variables for the dataset can provide novel information to the model and the value of each of these variables is directly tested by inclusion in the final model and relative contribution to the prediction algorithm.

DATA PARTITIONING

After finalizing the variables to be included into the model the next step is to partition the available data into training, revising and validation/testing data sets (Figure 1). Partitioning data into these data subsets prior to model building is necessary to evaluate predictive model performance and assure that the model can reasonably be applied to new datasets. The first data subset (training) is used to generate the initial models and this training data subset may differ from the native data related to frequency of outcome occurrence (see data balancing section below). The second data subset (revising)

is used as an initial test on the classification algorithms and predictive models created using the training data. The revising data may be processed through classification models multiple times with changes to model configuration or structure to optimize performance. Testing the model using training or revising subsets is inappropriate and could bias model accuracy upward. The third and final dataset is the validation or test data which is processed through the models only one time to generate predictive values for model assessment. Prior to the model building process, the data should be randomly partitioned into these three data subsets and the amount of data placed in each subset varies with project goals and frequency of target variable of interest. The hierarchical nature of the observations may influence the method of partitioning data and a structured randomized approach may be considered to divide observations among the data partitions. For example, if the dataset consists of multiple observations on individual animals, the individual animals may be randomly allocated to each dataset and each animal record would be tied to all of the accompanying observations. The method for partitioning the data are highly related to the overall classification goals and the underlying hierarchical data structure.

Training and revising data sets should be large enough to build and refine the model while retaining a representative sample of what future data might look like in the validation/testing data subset to accurately evaluate the performance of the model. Retaining both a revising data subset for model revisions and a validation/test data subset of data for final model evaluation is necessary. The revising data may be used multiple times in a single model or in multiple models with subsequent alteration of the model parameters based on these results, and if a separate testing/validation subset is not withheld, final model selection would then be based on the revising dataset which could lead to overfitting and less external applicability. There are no specific rules for the amount of data to partition into training, revising, and validation/testing datasets, but authors have successfully used partitions of 50% training, 25% revising, 25% validation/test (Abell et al., 2017) and 40% training, 30% revising, and 30% validation/test (Amrine et al., 2014). The specific size of the splits is based on data availability and frequency of outcome occurrence, but the data must be partitioned prior to the initiation of the predictive analytic process to avoid potential bias in final predictive assessment.

DATA BALANCING

The final consideration before building the predictive model is consideration of the frequency of outcome and if any adjustments need to be made to the dataset to account for the specific target outcome of interest. Rare outcomes (e.g., mortality) are often a target of predictive models because while infrequent in occurrence, mortality events can have significant economic impact. Several analytic tools can be valuable to generate models predicting rare outcomes, and modifications such as balancing the dataset can be important design considerations before building the model.

Balancing the training data set used to build the model consists of creating a training data subset with equal number of target events and nonevents, and using a balanced dataset has been shown to optimize performance of classification algorithms (Japkowicz, 2000). Balancing is used when the target outcome is rare and can be achieved by using oversampling or undersampling. Over-sampling relies on creating a training dataset by selecting all records belong to the minority class and creating duplicate records until the distribution of outcome is equal between events/nonevents. Undersampling is based on keeping all the minority class records and randomly removing records from the majority class until there are equal number of records in the target events and nonevents. Both methods can be effective, and one evaluation of mortality prediction in beef cattle illustrated some advantages of undersampling compared to the native, unbalanced dataset (Amrine et al., 2014). Models can be compared using the area under the receiver-operating characteristic curve to determine potential differences in overall classification accuracy or relative improvement over the natural occurrence rate of the target variable. Mortality predictions were created using the same series of classification algorithms (Bayesian network, decision stump, filtered classifier, boosted logistic regression, logistic regression, multi-boosted regression, naive Bayesian, random forest and voted perceptron) based on either a native dataset (prevalence of mortality = 8.5%) or a balanced dataset (prevalence of mortality = 50%). In this case, the median area under the curve for all models was greater when the dataset was balanced using undersampling compared to the same classification models based on the native dataset.

Only the training dataset is created in a balanced fashion and used to generate the predictive models. The revising and validation/testing subsets are left

in unbalanced fashion using the natural occurrence rate of events of interest; therefore, evaluations of model performance are based on native data that have not been adjusted using frequency of event occurrence. Balancing training data based on frequency of event occurrence can be an important consideration when generating predictive models for rare events.

MODEL SELECTION

Many different types of predictive models are available to classify data or generate predictive outcomes based on the specific data set. Big data include numerous variables that typically have complex interactions and relationships with the target outcome of interest; therefore, a predefined model selection based on the target variable is often not possible. The final assessment of which model provides the most valuable prediction is based on model performance when evaluating the accuracy of model predictions in the validation/testing data subset; therefore, multiple models are often employed to attempt prediction of the target outcome. By assessing the performance of multiple models, the evaluation does not assume data will conform to a specific form and the optimal model is determined not by preconceived notions, but rather by final accuracy evaluation. The predictive analytic framework provides the setting to test and evaluate multiple classifying algorithms to determine the best fit (based on accuracy) for a specific situation and target variable.

Classification methods vary in their ability to handle missing data, deal with different attribute types (continuous or categorical), their overall generalizability, and ability to provide a clear explanation of why a particular prediction was reached. The type of target variable and the level of prediction accuracy needed should help guide the specific algorithm chosen. Often, a good starting point is to build a simple classification model (e.g., a logistic regression model). This can provide a good baseline model and can be used as a point of reference when building more complex models.

Another good initial model framework is the decision tree due to its ability to be understood and be deployed in popular SQL-based database systems (Abbott, 2014). The decision tree framework can be used for both regression and classification problems. An example application is the use of a decision tree to evaluate data collected from accelerometers to classify animals based

on specific behavior types. In this research, the authors showed that some behaviors (lying/standing) could be accurately predicted based on accelerometer readings, whereas others (walking) had lower predictive accuracy (Robrt et al., 2009). Predictions from decision trees are made using a series of if-then-else rules applied recursively to the data, which result in more intuitively interpretable predictive models.

Ensemble methods combine the results of multiple models in an attempt to provide a more accurate predictive model compared to single-model methods. Random forests are a popular type of ensemble model that will often provide improvements in accuracy compared to single trees. In general, multiple trees are built trying to predict the same target variable using different combinations of predictors and subsets of the data set. These multiple models are then combined using various methods to determine the final predicted outcome.

One of the simplest models to understand predicting continuous outcomes is the k-nearest neighbor (**k-nn**) method. This model is easily described when using continuous variables to predict another continuous variable. The basic idea is there are a set of certain attributes that are associated with the target variable. For example, if height, weight, and ADG are known for a group of cattle, one could use those data to predict ADG on a new group of cattle based on their height and weight. By finding the cattle with the known ADG values that are closest in height and weight to an individual from the new group of cattle's height and weight, one can then assign or predict an ADG value. If three variables are used, this classification method could be imagined by picturing a 3-D plot of the data. Two dimensions are known on the target to be classified and this can be plotted to determine which other datapoints in the known space are the nearest neighbors. Then, using the attributes of the known nearest neighbors a prediction is created for the target variable. The "k" in k-nn represents the number of animals that are used after determining the distance measurement. The more neighbors used to create a prediction, the smoother the prediction; however, there is no theory that specifies the number of neighbors to choose (Abbott, 2014). Although the k-nn method is not as sophisticated as other methods discussed, depending on the target variable and predictor variables available, it can provide reasonably accurate results with minimal processing and is easily interpretable.

GENERATING PREDICTIVE MODELS

Once the preprocessing steps have been completed, training predictive models is the next step. Using the partitioned training data subset, individual algorithms are provided with the required parameters, the target variable is specified, and the model is trained. Several open source software packages such as R (R Core Team, <http://www.R-project.org/>, Vienna, Austria), Knime (Knime Analytics Platform, <https://www.knime.com>, Zurich Switzerland), RapidMiner (Rapidminer Inc, <https://rapidminer.com>, Boston, MA) have packages or nodes that provide different models that can be evaluated for classification problems. Many of these models have configurations or parameters that can be modified based on the specific situation. The initial configuration is based on experience with the models and similar type classification problems. After model generation using the training data subset, a predictive model is created and used to classify the revising data. At this point, the preliminary results are assessed evaluating how well the revising data were classified. An iterative process ensues allowing adjustments to the model, predicting results from the revising data subset, and assessing results. Model modifications during this process could include pruning of decision trees, inclusion or exclusion of variables, changing the number of iterations in a Bayesian process, or similar model configuration settings that may influence final classification accuracy. The type of classification errors (false positive/false negatives) are often not equally weighted in level of concern, and this revision process can be useful to insure classification errors are distributed in a manner consistent with the overall project objectives. This process can be completed several times allowing optimization of the model to the revising data subset. All adjustments should be made during this phase as the validation/test data subset will only be used once on the models for a final assessment of model performance.

Additional classes of classification models can be evaluated using this same framework. For example, a logistic regression, single decision tree, and a random forest might all be evaluated for their ability to classify a response. After the process of tuning for each type of model is accomplished, the validation/testing data subset would be used with each model and evaluation of the chosen metric (i.e., overall accuracy, sensitivity, specificity) would be compared among the models to determine the best predictive model. The final model selection process also includes factors such as model

interpretability, run time, and ease of deployment based on the expected use case.

ASSESSING PREDICTIVE MODELS

A critical component of the predictive analytic framework is assessing model performance and this is achieved by using the model created with the training and revising datasets to classify or predict target outcomes in the validation/testing data subset. Metrics to evaluate model performance vary based on the type of prediction problem. Overall, model accuracy should be evaluated and information on type of misclassifications can provide important information on future application of the model. Evaluation of model performance is based on a similar framework to evaluating a diagnostic test, or rather comparing the known true outcomes to model predicted outcomes (Figure 2). Internal model assessment includes comparison of the sensitivity and specificity. The sensitivity is based on the ratio of true positives and false negatives and provides an estimate of the ability to find positive cases. Model specificity compares the ratio of true negatives to false positives and provides an indication of the ability to find true negative cases. Because sensitivity and specificity are based on the true status of the predicted cases (the columns in Figure 2), they are not influenced by the overall prevalence or rate of occurrence of the event of interest. Thus, to truly evaluate the accuracy of the model, the positive and negative predictive values must be calculated. The positive predictive value represents the likelihood that a model predicted positive is truly positive and the negative predictive values represents the probability that a model predicted negative is truly negative. Even if the initial training dataset was balanced to account for rare occurrences, the validation dataset would include target event occurrence at the expected proportion within the population (native or unbalanced data). This allows the positive and negative predictive values to represent a gauge of expected model performance in field settings.

Evaluating all aspects of final model performance is necessary to get a true picture of how the model could be used in the field. For example, in a paper using a classifier to predict mounting behavior in bulls, one of the classifier models had a sensitivity of 86.6% and specificity of 77.8% (Abell et al., 2017). These are not unreasonable characteristics for a predictive test; however, the results must be considered in the context of the specific situation. Mounting events are exceedingly rare

		True status			
		Event	Non-event		
Predicted Status	Event	True Event	False Event	Apparent Event total	Positive Predictive Value
	Non-event	False Non-event	True Non-event	Apparent non-event total	Negative Predictive Value
		True Event total	True non-event total	Total population	
		Sensitivity	Specificity		

Figure 2. Representation of model assessment using results from classification of validation data with known outcomes (true status) and model predictions (predicted status) using the example of a classification variable with two potential outcomes (defined as event and nonevent).

when compared as a relative proportion of time (0.6% of data in this case). Thus, the negative predictive value of this model was 99.9% meaning that when the model predicted a mounting event was not occurring it was correct 99.9% of the time. Conversely, due to the rare nature of the predicted outcome, the positive predictive value was 2.6% or rather positive predictions were only correct 2.6% of the time. Evaluating the positive and negative predictive values of the model is critical to determine the final accuracy in a field environment.

LIMITATIONS OF PREDICTIVE MODELING

Big data can be used to generate a variety of predictive models with the goal of employing these predictive algorithms in future decisions. As the models were generated on a single dataset, this predictive analytic framework is susceptible to many of the challenges inherent in any retrospective data analysis including bias and limited external validity. The models are generated from a single data set; therefore, the utility of predictive models is directly linked to how representative the training, revising, and validation/test data subsets are of future data to be evaluated by the models. Bias is any factor that could systematically influence the outcome away from the truth (White et al., 2016), and big data do not inherently overcome bias. If data are biased when building and assessing the models, the outcomes and predictions will also be biased toward the initial dataset. There is not a statistical

or mathematical test for potential bias caused by variables that were not collected or included into the dataset. As predictive analytics are often performed on retrospective analysis, the scientist performing the evaluation is key in identifying potential sources of bias. Optimal outcomes from predictive analytic analysis of big data are contingent on utilizing an unbiased, representative data set to generate the classification models.

SUMMARY AND CONCLUSIONS

Big data in livestock production systems may be generated by operational data acquisition or through use of remote livestock monitoring technology. The predictive analytic framework can be applied in a systematic manner to create information from these data to enhance decision making for livestock production, health and welfare. Defining a target variable is similar to generating the hypothesis in live-animal experimental design and this variable influences the data management and type of models selected for analysis. Data partitioning allows a single dataset to be used for building, refining, and assessing model performance. The predictive analytic framework concludes with final model assessment and provides an evaluation of the accuracy of predictions including expected likelihood of finding events and nonevents as defined by the target variable. The predictive analytic framework provides a systematic methodology for evaluating big data to enhance livestock decision making and promote precision animal management.

LITERATURE CITED

- Abbott, D. 2014. *Applied predictive analytics: principles and techniques for the professional data analyst*. Indianapolis (IN): John Wiley & Sons.
- Abell, K. M., M. E. Theurer, R. L. Larson, B. J. White, D. K. Hardin, and R. F. Randle. 2017. Predicting bull behavior events in a multiple-sire pasture with classification algorithms. *Comput. Electron. Agri.* 136:221–227. doi:10.1016/j.compag.2017.01.030
- Amrine, D. E., B. J. White, and R. L. Larson. 2014. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. *Comput. Electron. Agri.* 105:9–19. doi:10.1016/j.compag.2014.04.009
- Grzesiak, W., P. Błaszczyk, and R. Lacroix. 2006. Methods of predicting milk yield in dairy cows: predictive capabilities of Wood's lactation curve and artificial neural networks (ANNs). *Comput. Electron. Agri.* 54:69–83. doi:10.1016/j.compag.2006.08.004
- Hayes, T., S. Usami, R. Jacobucci, and J. McArdle. 2015. Using classification and regression trees (CART) and random forests to analyze attrition: results from two simulations. *Psychol. Aging.* 30:911–29. doi:10.1037/pag0000046
- Japkowicz, N. 2000. Learning from imbalanced data sets: a comparison of various strategies. In: *AAAI 2000 Workshop on Learning from Imbalanced Data Sets*, Austin, TX. AAAI Technical Report WS-00-05, 10–15. Menlo Park (CA): AAAI Press.
- Larson, R. L., and B. J. White. 2015. First steps when using scientific literature in clinical veterinary practice. *J. Am. Vet. Med. Assoc.* 247:254–258. doi:10.2460/javma.247.3.254
- Robért, B., B. J. White, D. G. Renter, and R. L. Larson. 2009. Evaluation of three-dimensional accelerometers to monitor and classify behavior patterns in cattle. *Comput. Electron. Agri.* 67:80–84. doi:10.1016/j.compag.2009.03.002
- Theurer, M. E., Amrine, D. E., and B. J. White. 2013. Remote assessment of pain and wellness status in cattle. *Vet. Clin. North Am. Food Anim. Pract.* 29:59–74. doi:10.1016/j.cvfa.2012.11.011
- Theurer, M. E., D. G. Renter, and B. J. White. 2015. Using feedlot operational data to make valid conclusions for improving health management. *Vet. Clin. North Am. Food Anim. Pract.* 31:495–508. doi:10.1016/j.cvfa.2015.05.004
- Unruh, E. M., M. E. Theurer, B. J. White, R. L. Larson, J. S. Drouillard, and N. Schrag. 2017. Evaluation of infrared thermography as a diagnostic tool to predict heat stress events in feedlot cattle. *Am. J. Vet. Res.* 78:771–777. doi:10.2460/ajvr.78.7.771
- White, B. J., R. L. Larson, and M. E. Theurer. 2016. Interpreting statistics from published research to answer clinical and management questions. *J. Anim. Sci.* 94:4959–4971. doi:10.2527/jas.2016-0706
- Yoo, W., R. Mayberry, S. Bae, K. Singh, Q. Peter He, and J. W. Lillard. 2014. A study of effects of multicollinearity in the multivariable analysis. *Int. J. Appl. Sci. Technol.* 4:9–19. Available from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318006/>