Project Description: Animal Science & Big Data
Team: Luwei Lei, Ziqi Tang
To: Dr. John Yen
2/4/2019

Problem Descriptions

Dairy farming is one of the most important classes in agriculture. It is a study of long-term production of milk (*Solodev*). This subject is closely related to our daily lives. Milk is served as a basic component to other dairy products including butter, cheese, ice cream and so on. The United States of America is one of the largest producers of milk and milk products. According to the recent statistics by the United States Department of Agriculture (USDA), there were over 215,000 millions of pounds of milk production in 2017(*United States Department of Agriculture*). Behind this huge number, it is significant for farms to produce the milk smartly. Therefore, as Dr. Kevin Harvatine demonstrated in class, many researchers keep trying to investigate how to increase the efficiency of the milk industry in terms of quality such as milk fat, and the amount of production. They have discovered an inverse relationship between animal stress and milk productions. In addition, we can take a step further by utilizing the big data generated in dairy production. Dr. Harvatine provides us with 3 datasets, and from each, we will be able to gain some insights by applying big data techniques and knowledge. The first dataset, which is also the foundation of this project, is consisted with average genetic potential and last 6 to 12 observations of 8000 herd performance over the past year in the U.S. We will be able to investigate the relationship between genetic potential and herd production. The second dataset contains 1800 observations of the rumination, genetic potential, milk fat yield and milk fatty acid profile from five farms. This dataset will generate a deeper analysis of the importance of genetic potential. We could analyze the relationship between genetic potential and all other factors listed in the dataset. The third dataset will let us research on the additional problem that is provided by Dr. Harvatine. It contains about 2,200 observations from PSU trails with over 70 dietary treatments, which could be used to investigate the effect of cow and diet on milk fat yield and milk fatty acid profile.

Expected Significance of Models

The ultimate goal of our capstone project is to contribute to increase the efficiency of milk productions in farms by analyzing the affecting factors of cows via data science methods. Based on the data we have, we expect the predictive models to discover a reliable relationship between the genetic potential of cows and milk production, and possibly indicate the genetic potential producing the most amount of milk. In addition, we expect the models will find the correlations between genetic potential and other affecting factors such as rumination. The role of genetic potential is more important than people think. In 2017, a cow can produce almost 23,000 pounds of milk per year on average, whereas in 2008, a cow can only produce 20,500 pounds of milk per year. There's a 12% increase over the past 10 years (*United States Department of Agriculture*). Some elite herds can achieve 30,000-plus pounds per cow per year(*Brown*). If the farm could produce the same amount of milk with fewer cows, there will be lots of savings including, but not limited to, feeds, labors, electricity, and spaces. Therefore, our

project can discover how genetic potential can affect the production of milk and thus foster the cows with more beneficial genes, then achieve a long-term goal in producing milk efficiently.

Potential Insights from Models

Since the first two datasets both contain the information of genetic potential, we hope to gain some insights into the correlation between genetic potential and other factors. We will apply some statistical and machine learning methods to extract these insights, which will be explained in the later section. Besides systematic analysis, we will utilize the visualizations to explain our outcomes.

Planned Model Construction

There is "No Free Lunch". As Dr. Andrew Ng said, "Always begin by implementing a rough, dirty algorithm, and then iteratively refine it". We plan to start with linear regression and neural networks. The datasets we are going to deal with are numeric datasets. Linear regression allows us to fit a straight hyperplane to the dataset that is closest to all data points. It's the best way to discover linear relationships between the variables in the datasets. However, this model is not powerful enough to learn complex relationships. We do not know if this model is able to perform well in the datasets. Therefore, we will also implement neural networks, which is a extremely powerful method that can learn very complex relationships in the datasets.

For the neural network, we plan to start with a simple feed-forward backpropagation multilayer perceptron algorithm, which contains 4 layers(*Shahinfar*). This model is easier to implement and does not require too many computations, comparing to other deeper models. We will construct two networks in the first phase of the project: one for prediction of milk fat, another for prediction of the amount of milk. From there, we will decide whether to move on to a new prediction or to refine the model based on the performances of the models.

Planned Model Evaluation/Comparison:

We will hold out about 10% of the data records as the testing set and another 10% of the data records as the validation set. We then use these data to evaluate the performance of our models. If the left dataset is too small to train our neural networks, we will use a semi-supervised learning algorithm to expand the size of the dataset.

Milestones for Midterm Project Report:

Milestone 1: Finish data cleaning and feature engineering and building linear regression and four-layer network. We expect to discover the relationships between genetic features and the quality and the amount of the milk products.

Milestone 2: Refine the models if the performances of the models are poor. Move on to the next problem if the performances are promising. We expect the accuracy to be higher than 85% if possible.

Project plan:

2/11 - 2/17: Clean dataset & Feature engineering

- Clean up null values and outliers.
- Merge redundant columns if needed.
- Perform features correlation heatmap

2/18 - 3/3: An initial predictive model

- Finish implementing linear regression.
- Finish constructing the four-layer neural network.
- Fit the four-layer neural network to datasets, which may take time.

3/4 - 3/11: Refine the models

- Evaluate the models.
- Construct a more complex neural network if needed.
- Otherwise, move on to the next problem.
- Write up the midterm report.

3/12…: Plan after 3/12 will be finalized in the midterm report...

Reference:

Brown, Tim. “Are You Feeding to Meet the Cow's Genetic Potential?” *Progressive Dairyman*, www.progressivedairy.com/topics/feed-nutrition/are-you-feeding-to-meet-the-cow-s-genetic-potential.

Shahinfar, Saleh, et al. “Prediction of Breeding Values for Dairy Cattle Using Artificial Neural Networks and Neuro-Fuzzy Systems.” *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012, pp. 1–9., doi:10.1155/2012/127130.

Solodev. “Dairy Farming History | American Dairy Association North East.” *Www.americandairy.com*, American Dairy Association North East, www.americandairy.com/dairy-farms/dairy-farming.stml.

“United States Department of Agriculture.” *USDA - National Agricultural Statistics Service Homepage*, www.nass.usda.gov/Charts_and_Maps/Milk_Production_and_Milk_Cows/milkprod.php.

“United States Department of Agriculture.” *USDA - National Agricultural Statistics Service Homepage*, www.nass.usda.gov/Charts_and_Maps/Milk_Production_and_Milk_Cows/cowrates.php.