DS 440 Midterm Capstone Project Report - Team 3

Brian Nguyen - Computational Data Science
Jordan Reed - Computational Data Science

March 11, 2019

Contents

1 Project 1 Progress - Genetic Potential

Initially, our goal for project 1 was to use dataset 1 to analyze the relationship between genetic potential and milk, milk fat, and milk proteion production in cow herds. Through our analysis, our goal was to then build machine learning models to predict these production values through the given genetic potential features in dataset 1. If our analysis and models were to turn out to be successful, researchers could use this information to optimize cow genetics, and to predict production rates for herds. In turn, this could lead to better estimates of genetic potential in the future. Below, we discuss our work on project 1 so far, and any additions we plan to make in the future.

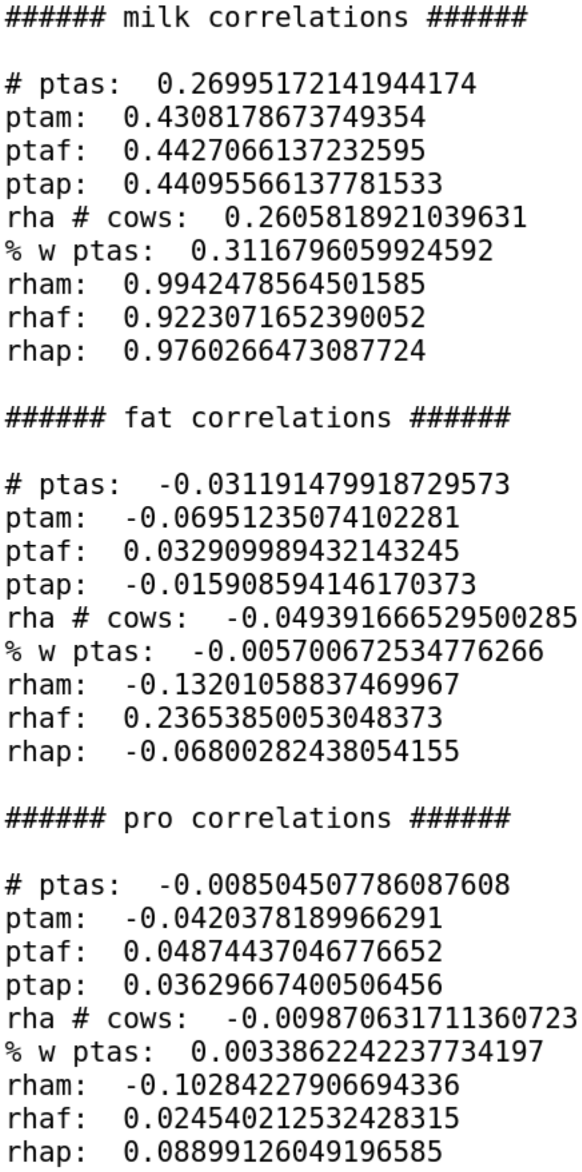
1.1 Data Preprocessing

Our work started off with performing preliminary cleaning of dataset 1. When we originally looked at the data, we noticed that much of the data had missing values in the form of spaces and dots, for example: " . ". We figured that dates with these values meant that observations were not recorded for those dates. This made sense since some of the herds didn't have values for dates 11 through 13. Sometimes it would occur that rhap and rhaf also had missing values. However, rhap, rhaf, and rham cannot simply be imputed with some value, since they are genetic features. Therefore, we decided to leave out samples if they did not have this inforamtion. For the other columns, we would either leave them as NaN or imputed the mean if they had missing values.

For each herd, we took the average of all available observations of herd performance and set those as our expected outputs of our machine learning models. We originally looked to use the month of the year as an additional input variable, but first had to test if the month-to-month difference was significant. As demonstrated in task 1, the standard deviation of the 6-12 observations for each herd individually is low on average - 0.2145 for fat, 1.7534 for milk, and 0.1610 for protein. So rather than worry about how or why the output changed depending on the date, we just took the average over those 6-12 observations and set that as our desired output. This allowed us to use a simpler model as our base model, such as the lasso regression model.

1.2 Feature Design

From our feature analysis we had discovered the Pearson correlations between each of our chosen features and the target variables (milk production, fat production, protein production) displayed below.



```
###### milk correlations ######

# ptas:  0.26995172141944174
ptam:  0.4308178673749354
ptaf:  0.4427066137232595
ptap:  0.44095566137781533
rha # cows:  0.2605818921039631
% w ptas:  0.3116796059924592
rham:  0.9942478564501585
rhaf:  0.9223071652390052
rhap:  0.9760266473087724

###### fat correlations ######

# ptas:  -0.031191479918729573
ptam:  -0.06951235074102281
ptaf:  0.032909989432143245
ptap:  -0.015908594146170373
rha # cows:  -0.049391666529500285
% w ptas:  -0.005700672534776266
rham:  -0.13201058837469967
rhaf:  0.23653850053048373
rhap:  -0.06800282438054155

###### pro correlations ######

# ptas:  -0.008504507786087608
ptam:  -0.0420378189966291
ptaf:  0.04874437046776652
ptap:  0.03629667400506456
rha # cows:  -0.009870631711360723
% w ptas:  0.0033862242237734197
rham:  -0.10284227906694336
rhaf:  0.024540212532428315
rhap:  0.08899126049196585
```

Figure 1: Correlations between Features and Target Variables

For milk production, each of the features displayed correlations ranging between 0.27 and 1. RHAM, RHAF, and RHAP in particular displayed very strong positive correlations close to values of 1. From plotting these three features against milk production, we can confirm their strong correlations, as shown below.

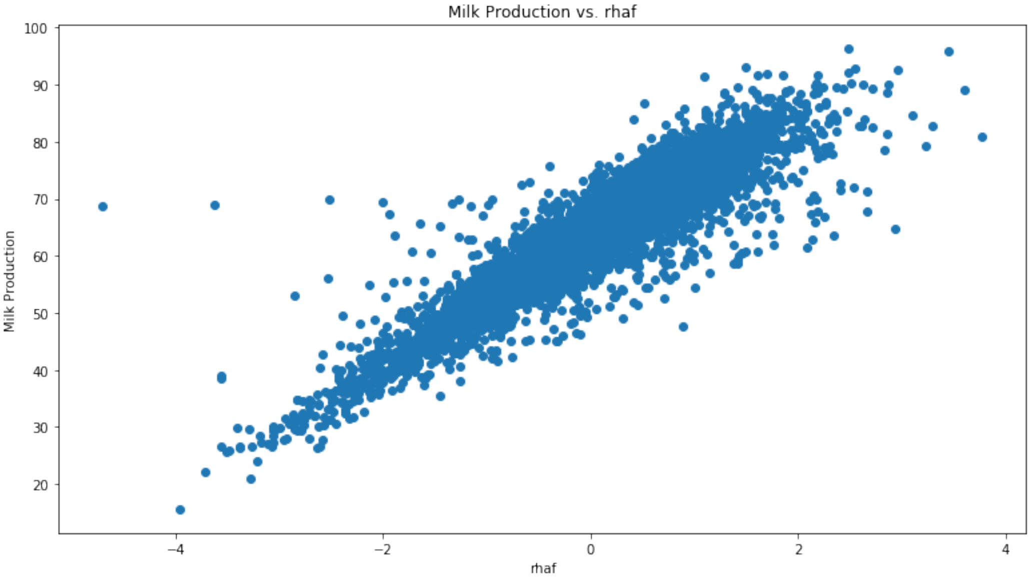


Figure 2: Relationship between rhaf and Milk Production

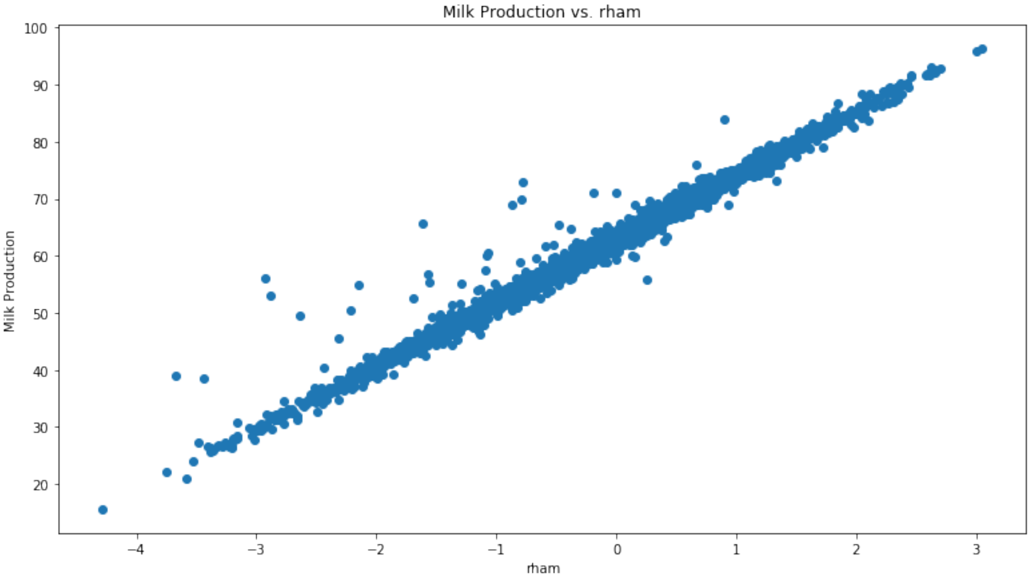


Figure 3: Relationship between rham and Milk Production

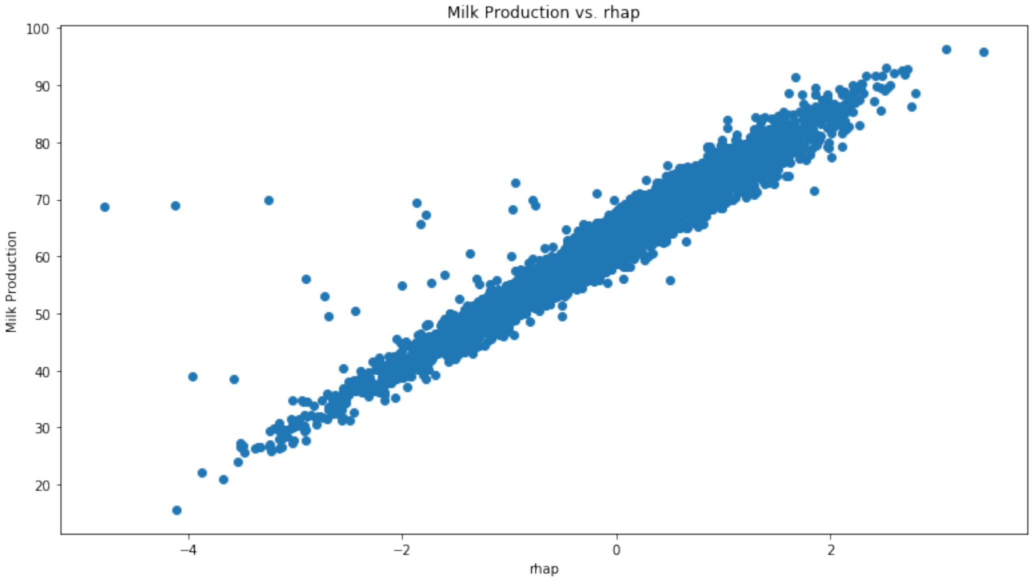


Figure 4: Relationship between rham and Milk Production

By knowing this, we altered the features we planned to use in our models for milk production to just be rhap, rhaf, and rham. We wanted to attempt to reduce the complexity of our milk production model by reducing the dimensional space to just those 3 columns. We tested this on our lasso regression model for milk production that we discuss below, and we noticed that the new adjusted-R^2 value was pretty much identical to the old one with all of the features. We figured that this could be explained by the fact that lasso regression is very good at getting weight vector values to 0 or close to 0, reducing the impact of features with low dependency. Even with the similar results, reducing the amount of features makes our models less complex and has potential of increasing speed for models down the line. The reduction of execution time will be tested with our random forest model.

The correlations for protein and fat production were underwhelming, since the most of the correlation values were close to 0. This was not too alarming due to our previous results. From our lasso regression model, we can assume that the combination of features with the learned weight vectors play a critical role in predictions, hence the high adjusted-R^2 values.

1.3 Lasso Regression

We started off using lasso regression. We ran three lasso regressions, using the same input parameters for all three - PTA's, PTAM, PTAF, PTAP, RHA cows, % W PTAS, RHAM, RHAF, RHAP. Each of them predicted a different type of production; the milk, fat, and protein production. We want to build models to predict these since we are studying the relationship between the genetic potential of herds and their productions. For these models, the process was automated by creating functions/loops to handle each production type. After normalizing the data by subtracting the mean and dividing by the standard deviation per column, the data was split into train 70% and test 30%. After a cross-validation with 10 folds for each, we were able to determine the optimal alpha value of 0 for each model. This is because the average of the R^2 values per production type for these folds were greatest at alpha=0. This also lets us use a less complex model. Due to this, our lasso regression model is actually more of an ordinary least squares regression. Results for the cross validation with varying regularization can be seen below.

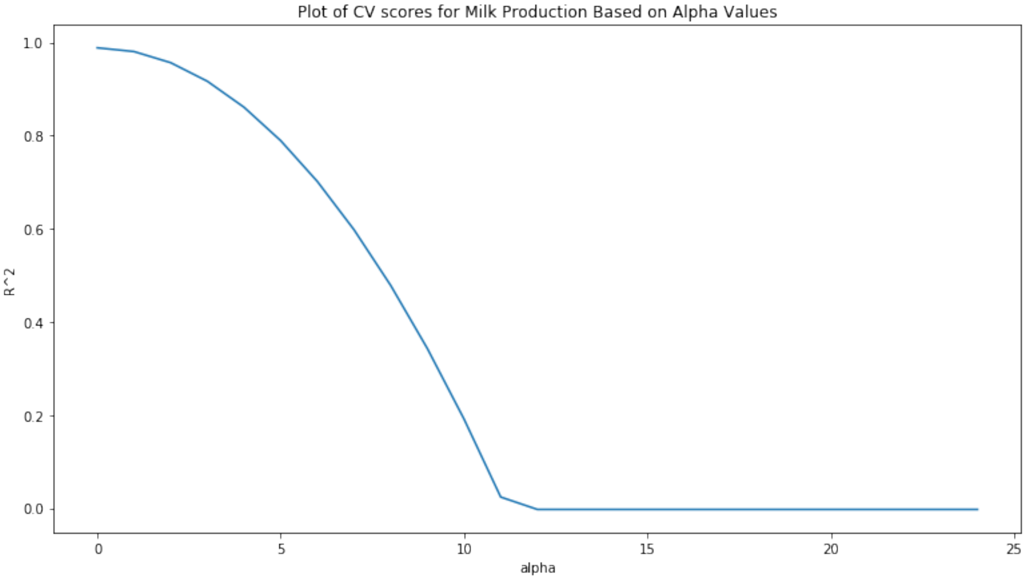


Figure 5: Cross Validation Results for Milk Production

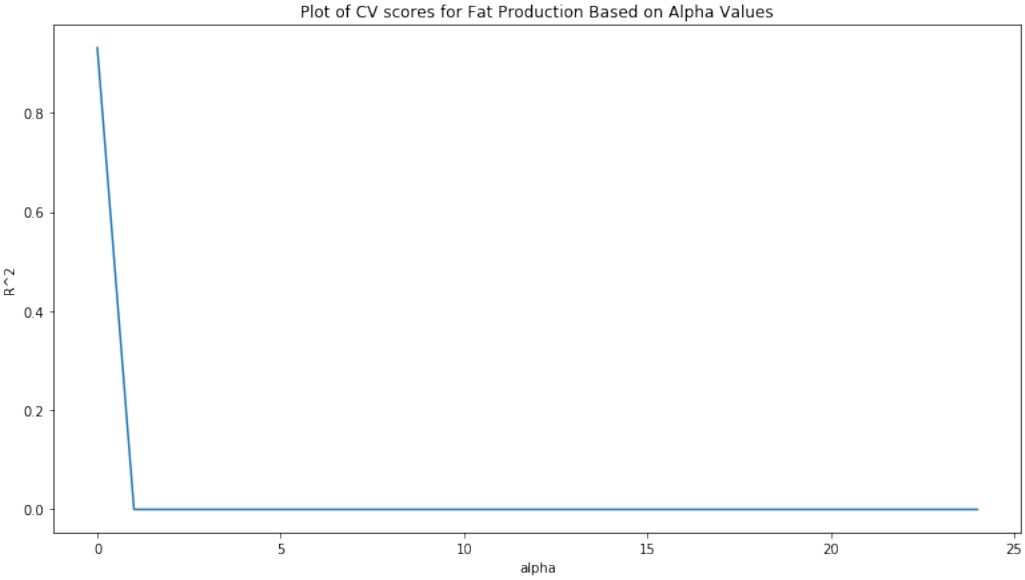


Figure 6: Cross Validation Results for Fat Production

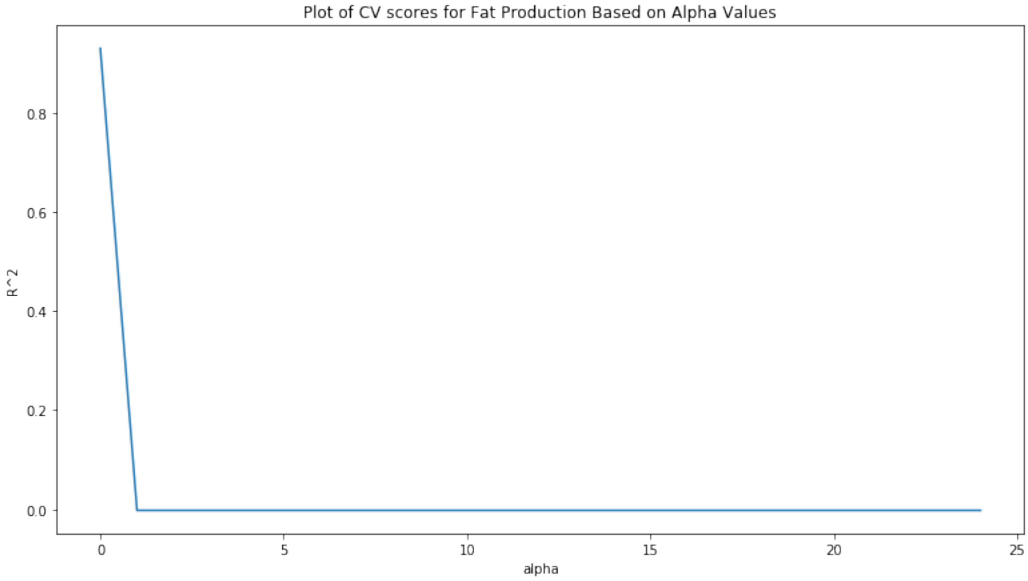


Figure 7: Cross Validation results for Protein Production

After predicting on the test set, we ended up with R^2 values of 98.92% for milk, 94.84% for fat, and 92.95% for protein. We also calculated the adjusted-R^2 value for our lasso regression model, and our results were very similar. This was just a way to confirm the reliability of our model, since normal R^2 has a tendency to fit better due to high dimensionality and tends to fit on random noise. The results can be seen below:

- milk: 0.9868
- fat: 0.9438
- protein: 0.9277

Lastly, we chose to include root mean squared error as one of our metrics because it is a great general purpose metric for showing how different predicted values are from actual values. It also increases error the higher the difference is in the results. This is important because it would alert us if some of our results are very off. Below you can see the RMSE results for each lasso model.

- milk: 1.3159
- fat: 0.0615
- protein: 0.0302

Considering the means for these target variables are about 55, 3.5, and 3 respectively, these results are very good. In addition to including RMSE, we wanted to visualize our prediction results against the truth values of our test splits for each model.

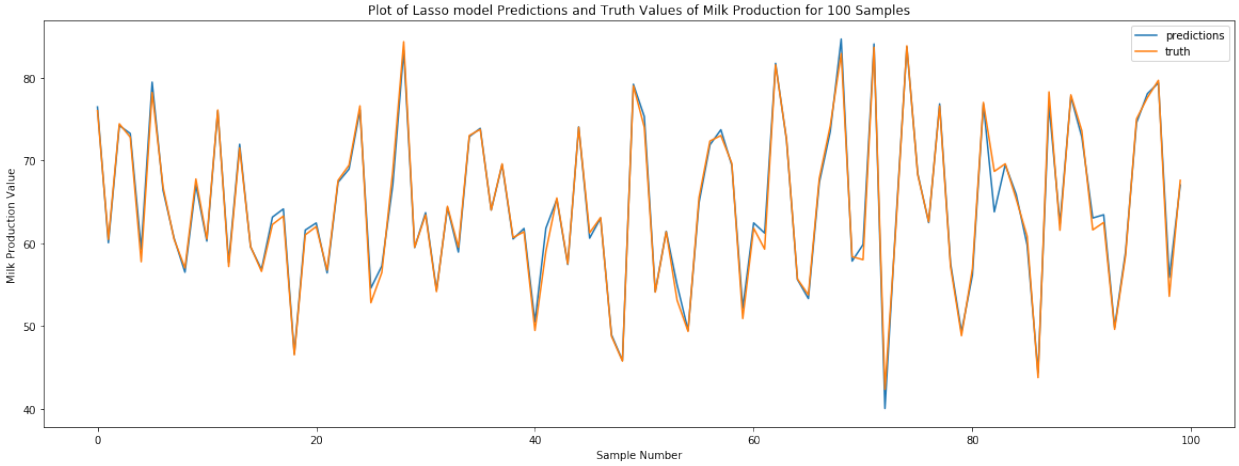


Figure 8: Plot of Lasso model Predictions and Truth Values of Milk Production

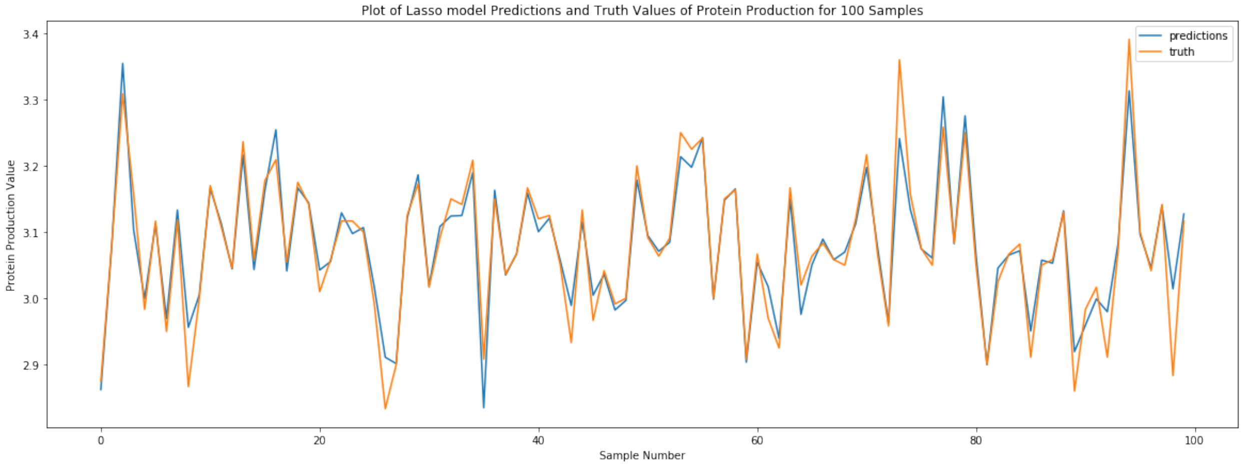


Figure 9: Plot of Lasso model Predictions and Truth Values of Protein Production

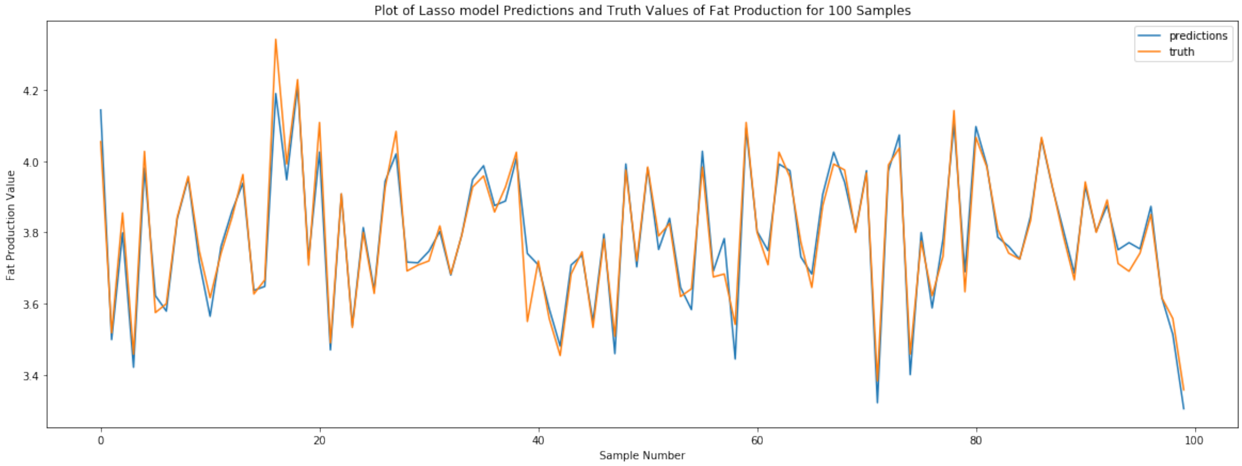


Figure 10: Plot of Lasso model Predictions and Truth Values of Fat Production

As you can see, predicted values are very similar to the truth values for the test set for each target variable. This confirms our evaluation metrics of adjusted-R^2 and RMSE.

1.4 Random Forest

We ran three random forest regressions to predict the exact same outputs, using the exact same inputs - PTA's, PTAM, PTAF, PTAP, RHA cows, % W PTAS, RHAM, RHAF, RHAP. Again, we normalized the data before running the regression. Our random forest model yielded high results similar to our lasso regression model, but not quite as high. Results for adjusted-R^2 are shown below:

- milk: 0.9831
- fat: 0.9365
- protein: 0.8963

1.5 Model Comparison

The execution time of the training and testing of the model was much longer than our Lasso Regression model. Random Forest with 1000 trees takes over 1 minute to execute for all 3 models while Lasso finishes in a matter of seconds. In real-world prediction and training, fast execution time is much more preferable, which would lead us to choose solely the Lasso model since it has both the better metric results, and much faster execution time. The only other downfall is that the random forest model could be more interpretable than the lasso regression model, as you can visually see what feature values produce what prediction.

We also tested the impact of our feature engineering on the random forest model for milk production. The execution time was a difference of 17 seconds between the two tests. If this were scaled to bigger data, we could see this playing a huge role in real-time training and prediction for cow milk production.

After the evaluation of the random forest model, we had decided not to continue the creation of our neural network. This is because our simple lasso regression model was performing well, and the more complex random forest model wasn't performing as well, so generating a more complex model that would take longer than lasso regression is frankly unnecessary.

1.6 Insights

RHAP, RHAM, and RHAF are rolling herd averages for each of their respective categories of protein, milk, and fat. This would be why they are the most telling features for our lasso regression models. Therefore, we believe that the actual use of these models in animal science wouldn't be of much use. This can be confirmed by the discussion with Dr. Harvatine, since his focus was mostly on datasets 2 and 3 of project path 1. We will be discussing more about our revised plans in later sections.

2 Project 2 Progress - Intake and Rumination Times

Our goal for project 2 was to use dataset 1 from project path 2 to see how we can predict intake and rumination times for cows. This would include features such as rumen composition, fermentation product, and plasma metabolites over a day for cows. Our analysis could show how voluntary cow intake and rumination times vary from cow to cow. One of the higher-level insights we could gain is if it is effective to optimize a farmer's time on an individual cow basis. If the time spent by the farmer performing these optimizations causes diminishing returns, we would determine that the predictions are not too applicable to real-world scenarios. Of course this would take into account the worth of an average farmer's time and how long it would take to do such optimizations. Additionally, following Dr. Harvatine's updated expectations, we are looking to predict rumen composition based on the aforementioned variables. The progress on project 2 that we have so far is discussed below.

2.1 Data Preprocessing

Our initial idea was to predict the amount of food consumed based on cow's blood contents, ammonia concentration, and volatile fatty acids content. After poking through the data, we decided to investigate an additional model that accounts for if a cow had been milked in the past two hours, with the idea that a recently milked cow will be hungry. An immediate concern is the lack of data - we had observations of each cow's blood, ammonia, and volatile fatty acids content for every three hours, and the intake values were for every half an hour. We didn't have much data to work with in the first place, so we decided to extrapolate data in between our three hour intervals for every input using a linear model. We initially considered just duplicating the rows, but there was far too much variation to consider that, and this extrapolation is naive enough as is.

Additionally, we had many potential output variables to consider. So far, we have only considered the raw output of food consumed each half hour, but we have many options to explore in the future. Following the conversation with Dr. Harvatine, it is clear that predicting rumen composition is a priority, and we have rumen data of each cow taken every three hours to work with. For the next task, we expect to develop models that predict these rumen compositions based on the inputs we've used for the models already developed.

2.2 Data Analysis

Before we dove into the model creation, we wanted to analyze if there was any clear relationship between the amount of food consumed and each of our potential input variables, so we knew where to focus our efforts. To this end, we computed the correlation between our inputs (each blood value, ammonia concentration, and total volatile fatty acids) and our expected output. The plots for BUN and Ammonia are depicted below.

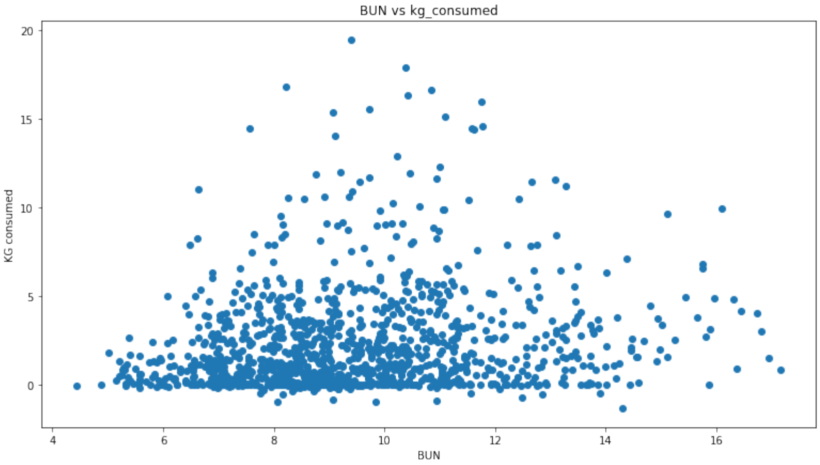


Figure 11: BUN vs KG consumed

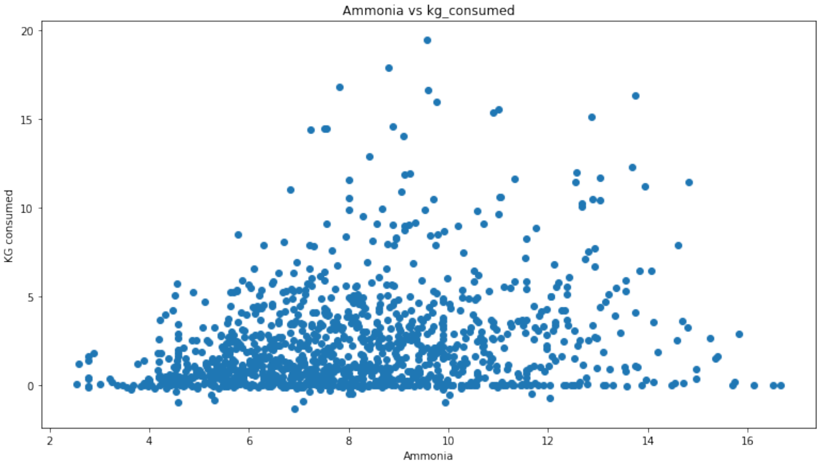


Figure 12: Ammonia vs KG Consumed

Those were the two highest correlations. As we can see, our correlations are not very good and negative for several features we were considering. As such, we only took the variables with correlation above 0.1 to consider in our model, to simplify our model and remove unnecessary data. This left us considering BUN, Ammonia, Acet, But, val, ButMR, and ValMR. However, But and ButMR are directly related to each other computationally, as are ValMR and val. As such, we dropped the MR version of these variables and considered exclusively But and val.

Another aspect to consider is the dominance of zero or near-zero values in our kg_consumed column. These are not very useful to predict - we care about how much a cow is eating at a certain time, not if the cow is not eating. This may have affected our models, so we expect to remove these values in the future to fine-tune our models.

Further, this also tells us that none of these variables are very related to the total amount of food consumed, whatsoever. We're going to go ahead and try to create the model, but we don't expect very high accuracy. This may lead to a shift in approach going forward - perhaps we'll attempt to predict a different variable, perhaps we'll try to predict the times directly, we have many options to consider. For now, we'll give this a shot

2.3 Model Creation

We went ahead and created a lasso regression and a random forest regressor for each of our two inputs - one with all the variables previously mentioned, and another with those variables plus a boolean indicating if the cow had been milked in the past two hours. We split the data into 70% training and 30% test, and went ahead and trained and evaluated our model. Our R^2 values are shown below:

- Without milked, lasso: 0.10965529485455561
- Without milked, random forest: 0.46800101102684216
- With milked, lasso: 0.08422622383397493
- With milked, random forest: 0.5474277231175181

As we expected, our models are really, really bad. We thought on it and realized that we weren't exactly answering the question at hand. Our model aims to predict the number of kilograms of feed a cow eats at a specific time of the day. The original question, as posed in Dr. Harvatine's lecture, was more along the lines of "At what times do cows eat?" We figured we could extrapolate the time eaten from the kilograms eaten every hour, but with our inaccuracy that won't be possible. From this point, we decided to let the regression sit for a while and shift our focus to a binary case - given a certain time, is the cow eating or not?

This problem is multi-fold. We first need to determine a threshold for if the cow is eating or not. Most of our values were above 0, but many were trivial, so we decided that all entries with <1 kg of food eaten would be considered the cow not eating, as they could be chalked up to error, food falling off the measuring device, etc. As such, we determined that about 56% of the time a cow is eating more than a kilogram of food per half hour, which is a reasonable benchmark for a potential model. Using this as our predictor, we went ahead and ran SVM and random forest classification using our base inputs - BUN, Ammonia, Acet, But, and val. The accuracy of our respective models is as follows:

- SVM: 0.6916666666666667
- Random Forest: 0.7555555555555556

The classifications are slightly better, particularly the random forest, but ultimately we're still missing something here. It's possible that we should be using more inputs (there's a lot more data on fatty acids and milk profile that we haven't used yet, for example), and we certainly need to optimize our parameters. In the future, we'll also consider using different ML models, and eventually stacking our models. We're also thinking about simplifying our models and considering solely Time, Sequence, and Treatment as inputs, using dummy variables for the categorical variables Sequence and Treatment. We have a lot more options to consider, so not all hope is lost yet.

3 Refined Project Plan

3.1 Future Work and Refinement

We plan to continue our work on project path 2 as planned, but updated the tasks per the discussion with Professor Harvatine. It was deemed that information for datasets 2 and 3 from project path 1 could be more useful to work on than dataset 1. As such, in preparation for this midterm report, we started looking into dataset 2, particularly how to predict two columns: AvgMilk and Fat%. Predicting these values using the data in dataset 2 is evidently the main goal of project path 1, as outlined by Dr. Harvatine, and as such we began investigating the best methods of doing so. Based on external research and some preliminary data analysis, many of the fatty acids presented in dataset 2 are likely not useful predictors, since they are produced by the cow and thus are only reliably measured alongside the values we want to predict. The preformed acids, however, come exclusively from food, and as such can be directly changed to influence a cow's milk output. We are currently working on developing these models, and will have our preliminary results available by the next task.

Additionally, for project path 2, we've made good progress on predicting if a cow is eating or not, but we need to expand our horizons with respect to this dataset. Predicting rumen composition is a lot more important than predicting intake times, it seems. The good news is, we have some data on rumen composition from this dataset, and we were told to use "any and all data available" to predict rumen composition. We will do so on our next task.

3.2 Risks

Since we could be working on project path 2 and on datasets 2 and 3 from project path 1, we may have too much on our hands. We will handle this by trying to re-align our vision with Dr. Yen and Professor Harvatine.

Also, our understanding of datasets 2 and 3 from project path 1 and the dataset from project path 2 may be weak, considering our results so far for our experiments are poor. This may be due to choosing the wrong features, or misinterpretation of the target variables. We will be dealing with this during our meeting with Dr. Yen this Friday, where we plan to ask questions regarding the datasets.

3.3 Project Plan

Task	Date/Time	Deliverables
Task 1	02/13 10:00PM	• Complete data analysis (pandas) for project 1 in a Jupyter Notebook Completed • Complete visualizations (matplotlib or bokeh) for project 1 in a Jupyter Notebook Completed • Perform any data cleaning on dataset 1 for project 1 Completed • Create initial lasso regression model for project 1 to predict milk, fat, and protein production, and evaluate it for feasibility / analyze weight vectors Completed
Task 2	02/20 10:00PM	• Complete any feature engineering, extraction, and selection on dataset 1 for project 1 Completed • Create other supervised machine learning models (Random Forest) for project 1 to predict milk, fat, and protein production Completed (neural network skipped)
Task 3	02/27 10:00PM	• Run a grid-search or randomized search for hyperparameter optimization for Lasso, Random Forest and NN Skipped, unnecessary • Generate RSME and R^2 metrics using 10-fold cross validation Completed • Complete tuning and evaluation for models for project 1 using cross-validation and a training-test split Completed • Perform any data cleaning on all three datasets for project 2 Completed • Complete data analysis (pandas) for all three datasets for project 2 in a Jupyter Notebook Completed • Create initial lasso regression models for project 2 to predict intake and rumination time (separately), and evaluate it for feasibility / analyze weight vectors Completed

Table 1: Table of Weekly Tasks (1)

Task 4	03/20 10:00PM	• Complete visualizations for project 2 dataset 1 (matplotlib or bokeh) in a Jupyter Notebook • Complete any feature engineering, extraction, and selection on all three datasets for project 2 • Clean data sets 2 and 3 for Project 1 Ongoing • Develop ML models to predict milk fat percent, total milk yield, fatty acid concentration considering the data in datasets 2 and 3 for Project 1 Ongoing
Task 5	03/27 10:00PM	• Create other supervised machine learning models (RF, Neural Networks) for project 2 to predict intake times • Create RF and lasso regression models for predicting rumination composition for Project 2 Ongoing • Evaluate our new project 1 datasets 2 and 3 models using 10-fold cross validation • Complete tuning and evaluation for dataset 2 and 3 models for Project 1
Task 6	04/03 10:00PM	• Run a grid-search or randomized search for hyperparameter optimization for Lasso, Random Forest, and NN • Generate RSME and R^2 metrics using 10-fold cross validation • Complete tuning and evaluation for models for project 2 using cross-validation and a training-test split
Task 7	04/10 10:00PM	• Create any needed visualizations for report • Finish final report and make last adjustments to code • Ensure all code is functional and works on cluster

Table 2: Table of Weekly Tasks (2)

4 Current Gantt Chart

Link to Chart