Predicting antimicrobial usage from farm data

Joe Mitchell

2022-10-29

# Github Link

<https://github.com/joebmitchell/C7081-Assignment>

# 1 Background

Antibiotic usage in the livestock industry has been a metric by which farmers have been increasingly measured and put under pressure to reduce (Hennessey et al, 2020). This has been mainly driven by the need to reduce livestock farming’s role in the development of antibiotic resistance as part of the One Health initiative. The main metric used for measuring the usage of antibiotics on farm is ‘mg/PCU’. This is a measure of mg of antibiotic used per standardised weight of animal on the farm. Calculating this metric can be difficult as working out the average weight of animals that has been on farm over the recording period is complicated to calculate correctly and collating the antibiotic usage is sometimes challenging.

A dataset was obtained from the Natural Environment Research Council’s Environment Data Service (Reyher et al. 2021). It contains farm management information and data on antibiotic usage for 53 dairy farms in South West England.

## 1.1 Objectives

1. Determine the key factors linked with a high mg/PCU on farm
2. Create a model which is accurately able to predict a farms antibiotic usage based on easily obtainable management factors

# 2 Methods

## 2.1 Data

The original data contained 4578 observations of 129 variables. These were composed of data from 53 individual farms in the South-West of England and the data was collected via a questionnaire, prescription records from veterinary practices and sample information taken during the study period. The data was initially stripped of all sample information and reduced to just farm level information including management and antibiotic usage information. This left 53 observations of 108 variables. The data was checked for missing values but none were present. There was one value miscoded as a “3” when “1” or “2” were the only possible values so it was changed to NA. The predictors were then reduced, leaving 28 predictors available for inclusion in the final model. The reasons for their removal are shown in Table 1.

Table 1 : Predictor reason for removal

| Number of Predictors Removed | Reason for removal |
| --- | --- |
| 21 | Sample information |
| 10 | Specific antibiotic usage |
| 16 | <5 Observations per factor level |
| 51 | Non-relevant predictors |
| 2 | Collinearity |

The final selection of variables available for selection in the models along with their coding is shown in Table 2.

The data was fitted to a variety of models using *total\_mg\_pcu* as the dependent variable and all other remaining variables as the predictors. The data was split into a training data set (n = 27) and a test data set (n = 26) in order to allow for the most objective assessment of predictive power (Barnard et al. 2019). These split data models were then used to calculate Mean Squared Error (MSE) in order to provide a value for the variance between the models predictions and the test data set. Although using a split data validation approach is considered the most objective assessment of predictive power cross validation was also used as to the small number of observations meant that the split of data may lead to some biased selections.

Table 2: Variable names, Description and Factor Levels

| Variable | Description | Factor Levels |
| --- | --- | --- |
| poultry | Have there been poultry on the farm during the project? | 1=yes, 0=no |
| equine | Have there been equids on the farm during this project? | 1=yes, 0=no |
| total\_cattle | Total number of cattle kept on farm (averaged over study, including males/beef calves) | 1=<250, 2=250-500, 3=>500 |
| yield | 305 day yield in litres (represents average milk yield per cow per year) | 1=<8000, 2=8000-9000, 3+>9000 |
| bought\_pre | How many head of cattle were bought in from elsewhere in the 12 months before the start of the project? |  |
| scc | Somatic cell count threshold for prophylactic dry cow therapy |  |
| pneum\_vacc | Are the heifers routinely vaccinated against pneumonia? | 1=yes, 0=no |
| diarrvacc | Are the heifers routinely vaccinated against diarrhoea, eg with Rotavec? | Y= yes, N=no |
| halocur | Is Halocur routinely used as a cryptosporidium preventative? | 1=yes, 0=no |
| trough\_clean | How often are water troughs cleaned for the pre weaned calves? | 1= daily, 2= less often |
| wean | When are calves usually weaned? | 1= less than 10 weeks, 2= greater than 10 weeks, 3=do on weight |
| pattern | Is the calving pattern seasonal or all year? | 1= seasonal, 2= most of year |
| hectares | How many hectares of land on the holding? Converted from acres by 0.4xacre where needed |  |
| bvdvacc | Are cows routinely vaccinated against BVD? | Y= yes, N=no |
| ibrvacc | Are cows routinely vaccinated against Infectious Bovine Rhinotracheitis? | Y= yes, N=no |
| leptovacc | Are cows routinely vaccinated against Leptospirosis? | Y= yes, N=no |
| lungvacc | Are cows routinely vaccinated against lungworm (eg with Huskvac)? | Y= yes, N=no |
| clostvacc | Are cows routinely vaccinated against clostridial disease (eg Blackleg/7 in 1)? | Y= yes, N=no |
| daystreatmast | How many days are routine mastitis cases treated for with the first choice treatment? |  |
| timesfirstmast | How many times a day is the first choice treatment for mastitis given? |  |
| injectmast | What percentage of clinical mastitis cases are given injectable antibiotics? |  |
| treatfoot | Do you routinely use a footbath to treat outbreaks of lameness? | 1= yes, 0= no |
| abcalve | Do you routinely use antibiotics following a difficult calving? | 1= yes, 0= no |
| acf | Do you have automatic cluster flushing? | 1= yes, 0= no |
| rat | How often are rats seen on the farm? | 1= occasionally, rarely or never, 2= more often |
| bedmilk\_2 | Is straw used in milking cow bedding? | yes= yes, no=no |
| water | Where does the drinking water supply for the farm come from? | 1= always mains, 0 = not always mains |
| total\_mg\_pcu | Total antibiotic usage in mg/ PCU |  |

## 2.2 Linear Regression

As the outcome variable is a continuous variable it was decided to use multiple linear regression. An initial multiple linear model was made using 30 variables (still included 2 variables removed for collinearity) and all observations. The GVIF values for the predictors were calculated and 2 variables removed as they were greater than 5.

The model was then remade using the final 28 predictors. The residuals for this model are shown in Figure 1 and show that there is significant variation in residuals especially at the extremes of the data.

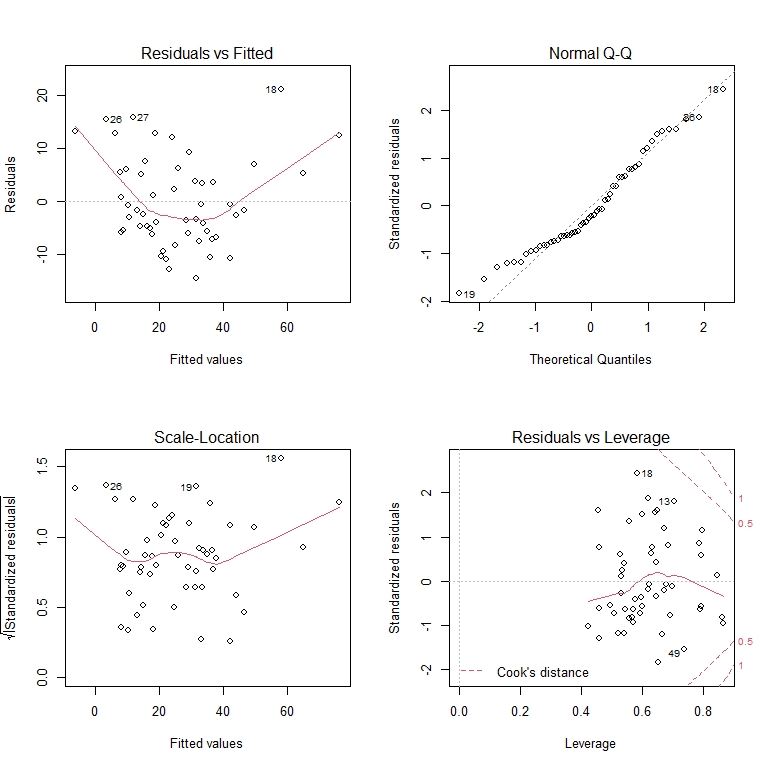


Figure 1: Residuals For Complete Linear Model

The GVIF values for the model were then calculated and are shown in Figure 2. Using a threshold of 5 only two predictors were above the threshold. The two predictors with GVIF of over 5 were factors with more than 1 degree of freedom. Their was compared to a threshold of (Fox and Weisberg, 2011). As they have 2 degrees of freedom they were both below the relevant threshold of 1.778.

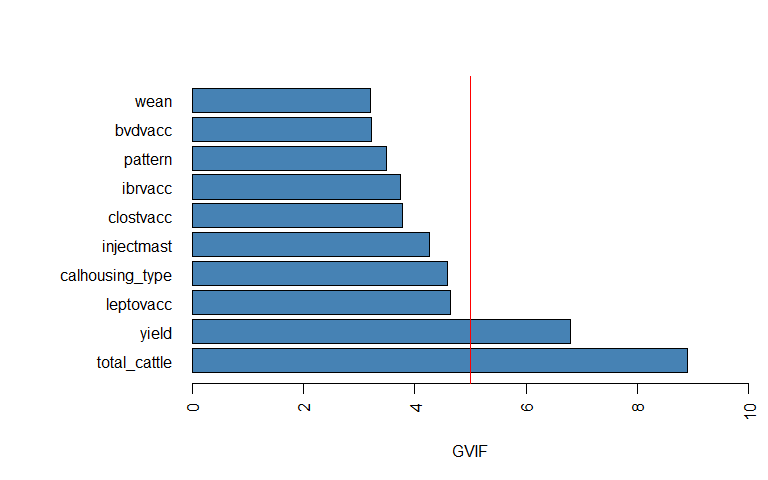


Figure 2: Top 10 GVIF Values

A multiple regression model was then created using the training dataset and predictions generated using the test data to calculate a MSE. As this split data now included less observations than predictors the model created had 0 degrees of freedom.

## 2.3 Subset Selection

Subset selection was then performed on the data in order to select the optimal subset of variables to include in a model. The regsubsets() function from the {leaps} package was used to perform the subset section. The three methods used were best subset, forward selection and backward selection. Test/Train data spliting was utilised for each subset selection as in the linear model. In addition Cross validation using k-fold cross validation was used for all 3 methods of subset selection using k= 10. This was performed as Xu and Goodacre (2018) have suggested that this can be a more accurate measure of performance in smaller datasets.

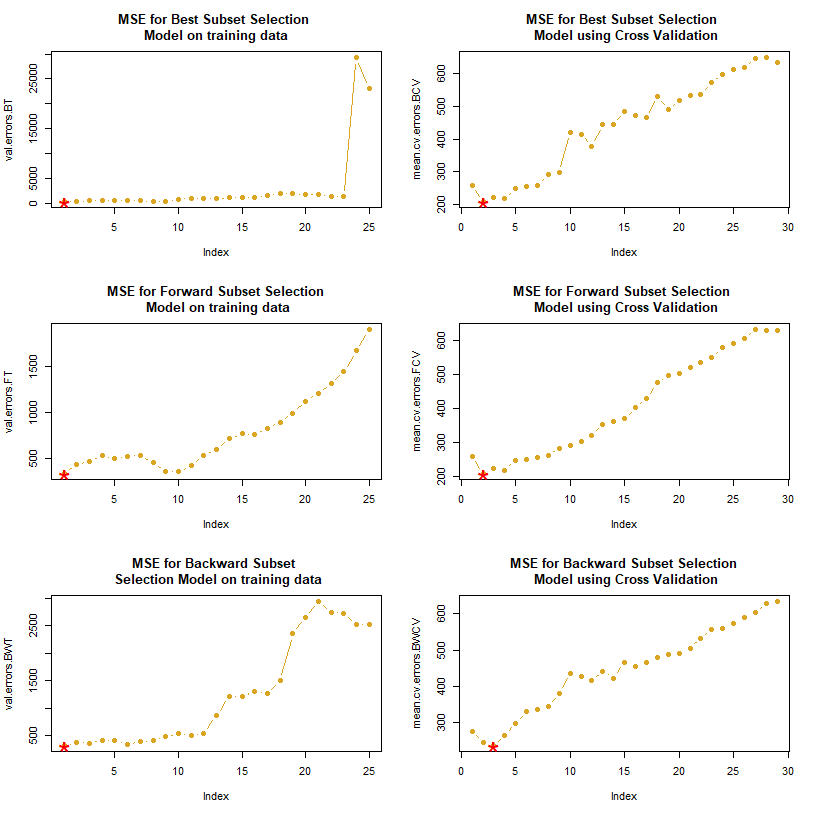


Figure 3: Impact of Model Size on MSE for subset selction

## 2.4 Shrinkage Methods

Ridge Regression was performed with the glmnet() fuction from the {glmnet} package. The strength of the penalty term is controlled by the tuning parameter and in order to select the optimal value cross validation was performed. The tuning parameter was set at 65.5.

Lasso uses L1 regularisation methods to add a penalty term to a regression model that has the effect of performing variable selection by forcing some of the coefficients to be exactly 0. The model was created with the glmnet() fuction from the {glmnet} package. Cross Validation was performed to select the value of lambda and a value of 1.4 was selected.

## 2.5 Component Reduction

Principle component regression is a method that applies Principle Component analysis to the data in order to perform dimension reduction in an unsupervised manor and then use the output as new regressors. The regression was performed using the pcr() function from the {pls} package. Cross Validation was performed to select the optimum number of components to include (M=5).

Partial Least Squares is very similar to PCR in that it attempts to perform dimension reduction and then uses the output as the new regressors but the main difference is that the transformation is supervised meaning that if the directions with low variance have a high predictive power then they will not be dropped unlike with PCR. The regression was performed using the pcr() function from the {pls} package. Cross validation was performed to select the optimum number of components to include (M=2).

## 2.6 Tree Based

Tree based methods were used to create further models. The trees were all trained on the training dataset and then predictions made on the test data to measure their performance. A simple regression tree was built first using the tree() function from the {tree} package. Pruning was then performed using cross validation in order to limit overfitting. In this model cross validation selected 4 terminal nodes.

Bagging was then used to try and improve the accuracy of the regression tree as it can reduce the variance of the predictors. This was performed with the randomForest() function from the {randomForest} package. As bagging improved the predictive accuracy of the model, Random Forest modeling using the same function was also performed using the default value for mtry of p/3 which worked out at 8. Boosting was performed using the gbm() function in the {gbm} package. n.trees was set to 5000 and the interaction depth to 6. Finally a Bayesian additive tree was created using the gbart() function from the {BART} package.

# 3 Results

The calculated MSE for the above models are shown in Figure 4. The lowest MSE was found in the best subsets and forward subsets when measured using cross validation. This however is an unfair comparison as the other models were all validated using test/train data. Bagging resulted in the lowest MSE for a model validated with test/train data at 232.1. A MSE of this magnitude tells us that the average prediction was incorrect by +/- 15.2mg/PCU which is quite a high margin when the target is <21mg/PCU.

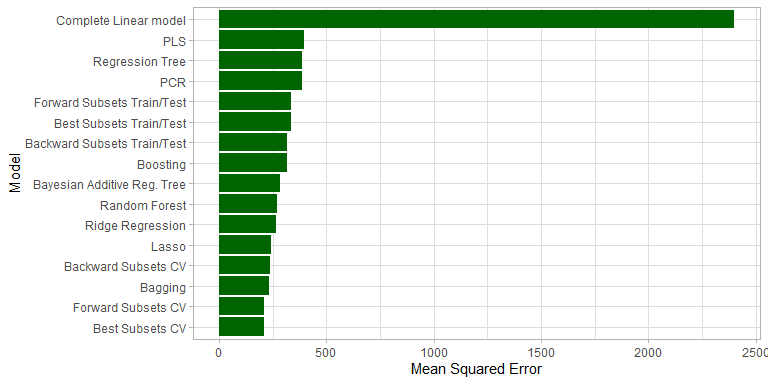


Figure 4: MSE for all models

The bagging models %IncMSE and IncNodePurity are shown in Figure 5. Using the percentage increase in MSE figures the 5 most important variables for the model are *daystreatmast*, *bvdvacc*, *injectmast*, *halocur* and *diarrvacc*.

It is to be expected that variables such as *daystreatmast* and *injectmast* would affect a farms antimicrobial usage as they will directly increase a farms usage, and if we had information on a farms mastitis incidence it would be possible that the interaction between these variables would increase their predictive power.

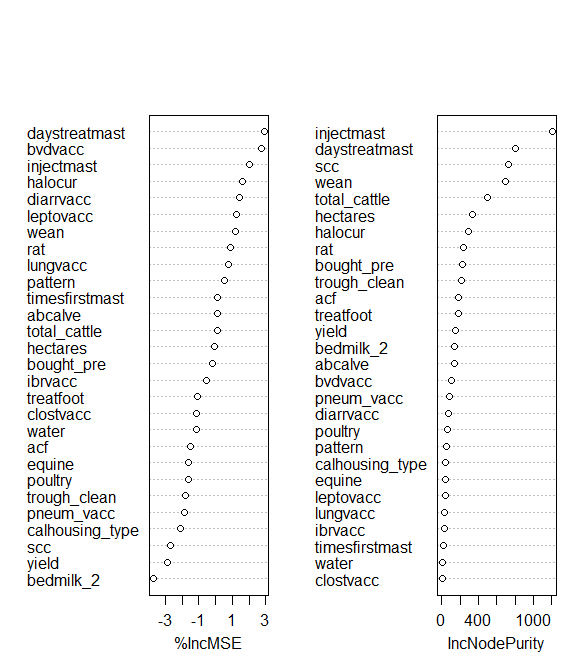


Figure 5 : Bagging model importance

The predictions made using this model compared to a farms actual are shown in Figure 6. This graph confirms that although there is reasonable predictions generated by the model particularly at high actual mgPCU the model is quite inaccurate.

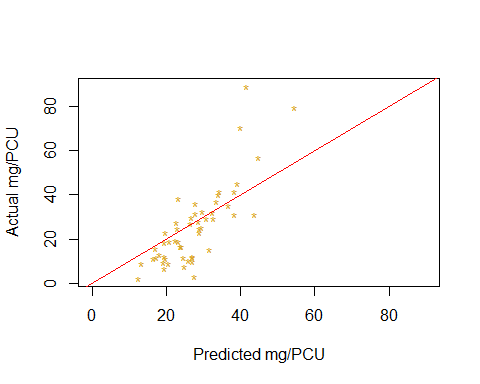


Figure 6 : Bagging Model’s mg/PCU predicted vs actual for complete dataset

Although bagging produces the lowest MSE and so provides the prediction with the highest accuracy (Breiman, 1996) it can be argued that it a more useful model for our first outcome is our lasso model as by performing variable selection and reducing some coefficients to 0 it leaves us with a smaller list of coefficients which are important when determining a farms antimicrobial usage. The variables coefficents are shown in table 3 and the largest coefficients and so the biggest effect on the prediction were *wean2*, *daystreatmast*, *treatfoot1*, *bvdvaccY* and *wean3*.

Table 3: Lasso model coefficients

| Variable | Coefficient |
| --- | --- |
| wean2 | -9.3989759 |
| daystreatmast | 4.1476456 |
| treatfoot1 | 3.3260229 |
| (Intercept) | 3.0180019 |
| bvdvaccY | -2.6385962 |
| wean3 | -2.5608854 |
| trough\_clean2 | 2.3307836 |
| timesfirstmast2 | 1.8094374 |
| halocur1 | 1.6988414 |
| yield3 | 1.5224282 |
| yield2 | -0.2411614 |
| injectmast | 0.1927635 |
| bought\_pre | 0.0421808 |
| poultry1 | 0.0000000 |
| equine1 | 0.0000000 |
| total\_cattle2 | 0.0000000 |
| total\_cattle3 | 0.0000000 |
| scc | 0.0000000 |
| pneum\_vacc | 0.0000000 |
| diarrvaccY | 0.0000000 |
| pattern2 | 0.0000000 |
| calhousing\_type2 | 0.0000000 |
| calhousing\_type3 | 0.0000000 |
| hectares | 0.0000000 |
| ibrvaccY | 0.0000000 |
| leptovaccY | 0.0000000 |
| lungvaccY | 0.0000000 |
| clostvaccY | 0.0000000 |

Its predictions are shown in Figure 7 and it is follows a very similar pattern to the bagging model where it is relatively accurate at lower actual mg/PCU but worsens as the values increase. The data in this area is relatively sparse as very few farms had a mg/PCU of >50 and so it is difficult for the model to accurately model these,

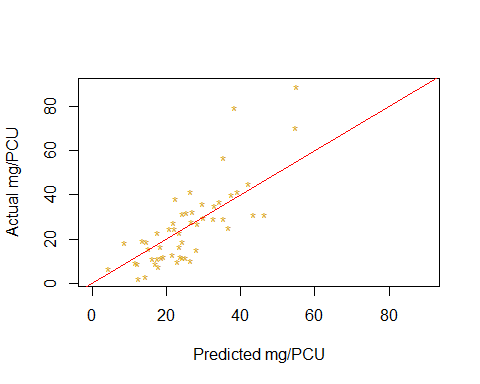


Figure 7: Lasso Model’s mg/PCU predicted vs actual for test dataset

# 4 Conclusions

The factors on farm that have been identified as important predictors of a farms antimicrobial usage in the lasso model and the bagging model include weaning age, treatment length of mastitis and if they inject mastitis cases with antibiotics, halocur usage, diarrhoea and BVD vaccination, and if they use a footbath to control lameness. These factors could be used when talking to farmer to identify farms that may require a more indepth discussion about their antimicobial usage and perhaps calculating their antimicrobial usage. Of course these factors are not all likely to be directly causing an increased usage of antibiotics but may just be correlated with other factors that increase usage.

The second objective of creating a model that could accurately predict a farms antimicrobial usage. The best performing model was still not very accurate and so would not be recommended for estimating a farms mg/PCU although it could be a useful starting point for engaging a farmer in a conversation based on a few simple questions rather than an in depth analysis of his figures. To further validate the model a larger nationwide sample of farms could be tested using the model. This would test its ability to predict usage on a wider sample of farms and provide a valuable validation of the model.

# 5 References

Barnard, D.M., Germino, M.J., Pilliod, D.S., Arkle, R.S., Applestein, C., Davidson, B.E. and Fisk, M.R. (2019). Cannot see the random forest for the decision trees: selecting predictive models for restoration ecology. Restoration Ecology, 27(5), pp.1053–1063. <doi:10.1111/rec.12938>.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), pp.123–140. <doi:10.1007/bf00058655>.

Fox, J. and Weisberg, S. (2011). An R companion to applied regression. Los Angeles: Sage.

Hennessey, M., Whatford, L., Payne-Gifford, S., Johnson, K.F., Van Winden, S., Barling, D. and Häsler, B. (2020). Antimicrobial & antiparasitic use and resistance in British sheep and cattle: a systematic review. Preventive Veterinary Medicine, 185, p.105174. <doi:10.1016/j.prevetmed.2020.105174>.

Reyher, K.; Avison, M.; Schubert, H.; Cogan, T.; Gould, V. C. (2021). Farm management and longitudinal data on antibiotic use and antibiotic resistant E. coli for 53 dairy farms, South West England, 2017-2019. NERC EDS Environmental Information Data Centre. <https://doi.org/10.5285/c9bc537a-d1c5-43a0-b146-42c25d4e8160>

Xu, Y. and Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing, 2(3), pp.249–262. <doi:10.1007/s41664-018-0068-2>.