Poultry Vocalization and Mortality

David Howell
&
Team

Contents

Introduction

As the human population continues to grow, so does the need for food and protein. Chicken is the most energy-efficient of all major meat categories in the US. Poultry production has a 13% caloric conversion efficiency[1], and a 21% protein conversion efficiency[2], which is superior to pork (9% caloric, 9% protein), and beef (3% caloric, 3% protein) in both measures. Poultry farming is also the most space-efficient meat production in the United States [12]. Shepon et al. estimate that substituting beef with poultry in the American diet has the potential to feed an additional 116 to 142 million people through increased feed conversion and the reallocation of pasture land for other agricultural purposes. Poultry farming is the least environmentally damaging form of meat production, and increased efficiency in the poultry industry through automated monitoring of chicken welfare has the potential to make a large positive environmental impact.

Poultry farms operate on razor thin margins, and the health and welfare of poultry stock makes a significant impact on profits. Consumers have also demonstrated an economic preference for animal welfare, and negative media coverage of “factory farming” practices have a small negative effect on chicken consumption [13]. An increase in animal welfare lowers costs and increases demand, providing economic benefits to poultry farmers.

Company Sponsor

Our project sponsor was a company (‘The Company’) that uses machine learning and signal processing to detect sickness in poultry farms and monitor overall chicken welfare. The Company aims to increase the health of chicken flocks and decrease farm labor costs through automated live audio and video monitoring.

Poultry Mortality

At large-scale poultry farms, tens of thousands of chickens are raised from early life to slaughter weight in large coops. With space limited, animal fitness is low, and diseases and injuries proliferate. Each day, farmers perform ‘culls’ and remove any dead, seriously injured, or obviously sick birds from the flock. The number of birds that are removed each day is recorded as the flock’s daily mortality.

The mortality rate of a flock is highly important to a poultry farmer. The obvious reason is that the farmer is unable to sell any birds that are culled from the flock, thus decreasing revenue. High mortality also has a negative economic impact on poultry farms because any food, water, and other costs spent on culled animals become sunk costs. The impact of these costs is greatest when mid- or late-life animals are culled, as by that point there has been a large amount spent on their development.

---

[1] Human consumed calories / total feed calories
[2] Human consumed protein / total feed protein

Mid- and Late-Stage Vocalizations

The Company tracks each flock through a full development cycle. Birds are considered mid-stage around three weeks in age. Late-stage begins at five weeks. As birds age, they learn a greater range of vocalizations. Mature chickens use vocalizations to communicate information about their surroundings, the discovery of food, pleasure, distress, and more. Our group focused on the three most common distinct chicken vocalizations: clucks, trills, and squawks. Trills are mid-to-high pitched vibratory sounds that last close to a second and indicate happiness. Clucks are lower-pitched short noises. Squawks are high-pitched, sharp noises that last around a second, and may indicate some form of distress.

Problem Statement

Much of the Company's work to-date has been focused on monitoring early-life chickens, and classifying vocalizations indicating obvious sickness, e.g. coughs. Mid- and late-life vocalizations have not been as extensively studied, and developing an effective classifier for them can facilitate further research on this period of the birds' lives. The factors that contribute to chicken mortality have also not been studied in-depth as part of the practicum program, and mortality mitigation can provide significant economic benefits to poultry farmers. Our goals for this project were twofold:

1. To develop techniques and models to effectively identify common mid- and late-life chicken vocalizations
2. To explore the relationship between various factors (including vocalization frequencies) and chicken mortality

Data

Raw Data

Our raw data consists of audio recordings taken from inside a large-scale chicken coop. Several microphones are located throughout each chicken coop, and continuously record audio data. The audio data is split up into minute-long FLAC files and uploaded into Amazon S3. The placement of the microphone has a significant impact on the quality of audio data for the purpose of vocalization classification. The Company microphones are placed along water lines and feed lines in the chicken coops. Microphones near feed lines pick up more frequent scratching and pecking sounds, as birds tap their beaks while feeding. Therefore, we picked a microphone located on a water line, in order to minimize the amount of undesired sound.

Feature Engineering

To convert raw audio data into usable features, each FLAC file is first converted to an audio signal, which is a waveform that measures the change in atmospheric pressure over time. The signal is converted to a spectrogram by applying the Fourier transform to overlapping windowed segments of the signal [11]. The spectrogram is converted to the Mel scale, which is based on human perception of frequency. Finally, n features are calculated by applying n triangular filters to the Mel spectrogram. The log-energies captured by the filters formed the basic features used to train our models. The code used to engineer the log-Mel energy features was provided to us by the Company.

Noise Reduction

One difficulty of working with audio data from mid- and late-life is an increase in fan noise. As chickens grow, they begin producing large amounts of body heat and waste. To dissipate heat and moisture in a confined space, chicken farms use large fans, which blow air into the houses near-constantly in the late stages of development. This persistent background noise presents a large obstacle for audio classification.

To reduce the effect of fan noise, a Weiner filter can be applied to the audio signal before conversion into the Mel-spectrogram. The Weiner filter attempts to isolate an underlying signal by decomposing an input signal into the desired signal (chicken vocalization) and additive noise (fan noise). All of the audio samples were run through a noise reduction algorithm provided by the Company, and the noise reduced samples were used for labeling, as they facilitated identifying chicken vocalizations by ear. Figure 1 shows the unprocessed signal from one of our labeled audio files, and Figure 2 shows the signal after the noise reduction algorithm was applied.

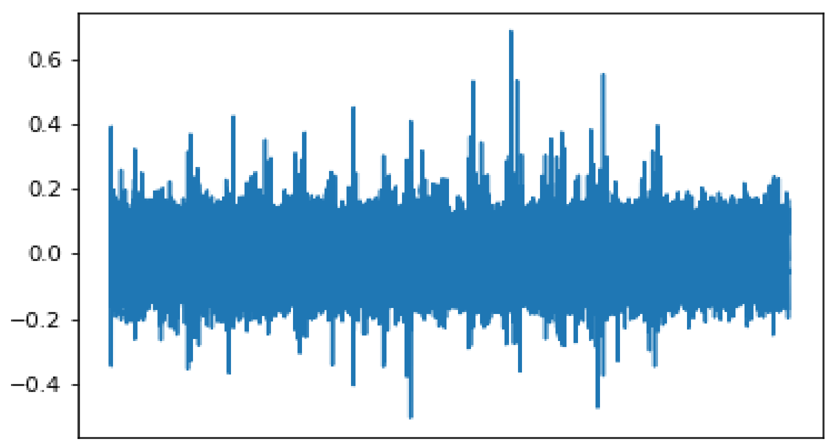


Figure 1 — Unprocessed audio signal

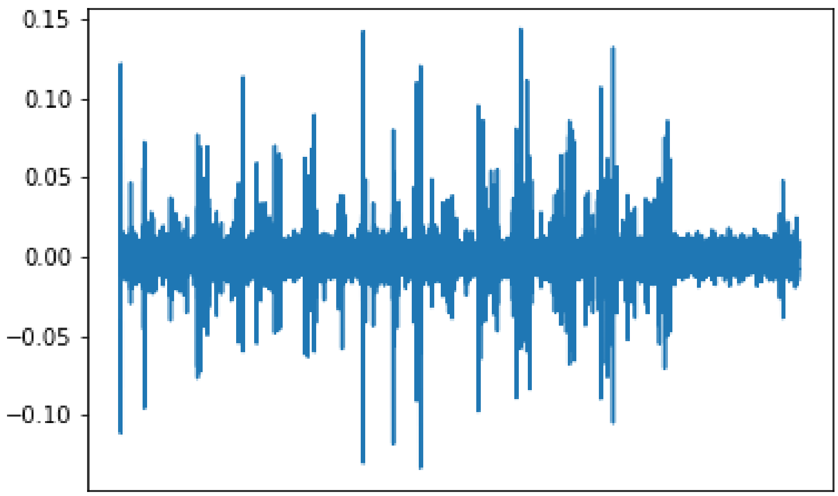


Figure 2 — Noise-reduced audio signal

Labeling

To generate ground-truth labels, we first coordinated a taxonomy of chicken vocalizations. Our team selected a handful of files and individually labeled each of them. We compared our individual labels until we were satisfied that we had created a consistent labeling scheme. We then randomly selected audio files from the beginning of mid-life until the end of the flock. We chose a total of 152 audio files and manually labeled all clear instances of “cluck”, “trill”, and “squawk” using the audio editor Audacity. The majority of the audio does not contain chicken vocalizations, so our dataset contains a large class imbalance, with the majority class being ‘no call’, or audio without a distinct chicken vocalization present. The final count of labels in our dataset was as follows:

- no call: 359,377
- cluck: 3,780
- trill: 543
- squawk: 1,117.

Vocalization Classification

XGBoost Vocalization Classification

Hyperparameter Selection

XGBoost is a gradient-boosted decision-tree-based algorithm [14]. I randomly split the data into training (70%), validation (15%), and test (15%) sets using a consistent seed. For each model iteration, I used Sagemaker’s Bayesian hyperparameter tuning, which balances exploration and exploitation to find a set of hyperparameters that best fit a validation set. The best performing model was then used to predict labels on the test set, resulting in the metrics reported below.

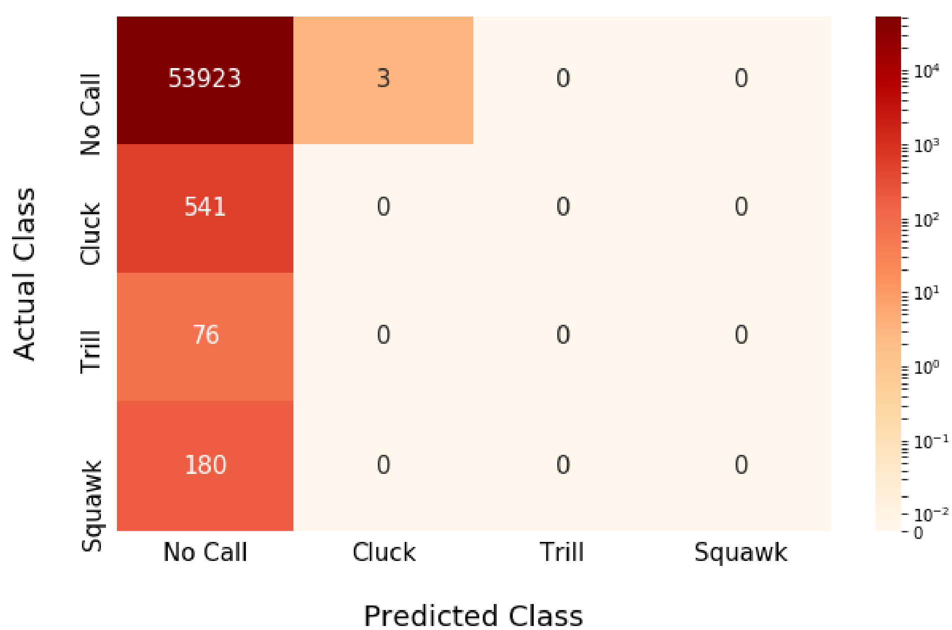
Model Evaluation

The primary metrics I used to compare models were accuracy, which is the percentage of samples that are classified correctly, and macro-weighted F1-score. The F1-score is the harmonic mean of precision and recall. The macro-weighted F1-score is the simple average of the F1-scores for each class. All classes are given equal weight for this measure.

Basic Model

I focused on building a multiclass classification model to identify clucks, trills, and squawks using XGBoost. My first iteration of the model was built using the log-Mel energy features. My initial model showed good accuracy, classifying 98.6% of test audio samples correctly. Unfortunately, that was because my model only classified three samples as clucks, and classified the rest as 'no call'. This was a result of the large class imbalance in the training set. The macro F1-score for the model was a disappointing 0.248. A model that predicts only one class when many are present is of little utility.

Table 1: Confusion matrix for basic model

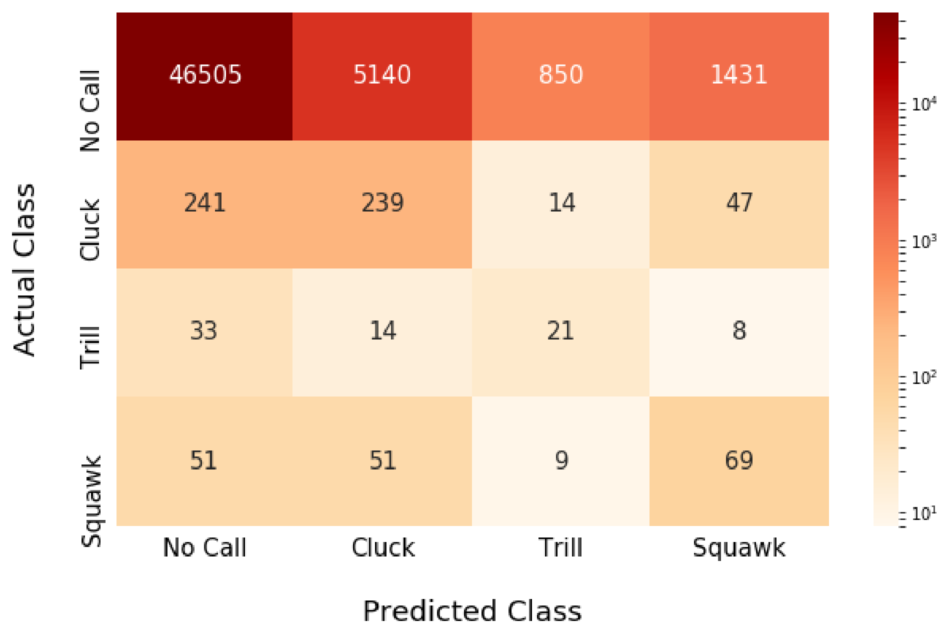


Upsampling

To mitigate the class imbalance problem in our dataset, I created a function to upsample the training data. For each minority class, it randomly duplicates samples with repetition to bring the minority class to the same size as the majority class. After upsampling, the training set contains greater than one million samples. The test and validation sets are not upsampled.

The model trained on unsampled data predicts many more vocalizations, but the vast majority are false positives. Upsampling alone creates a much less accurate model, resulting in a test accuracy of 85.6%, compared to 98.6% from the initial model. The macro-weighted F1-score of the upsampled model was 0.281, a slight improvement over the model without upsampling. I trained models without upsampling for each of the variable combinations mentioned below, but all of them had the same core issue: they classified everything as ‘no call’.

Table 2: Confusion matrix for upsampled basic model



Deltas

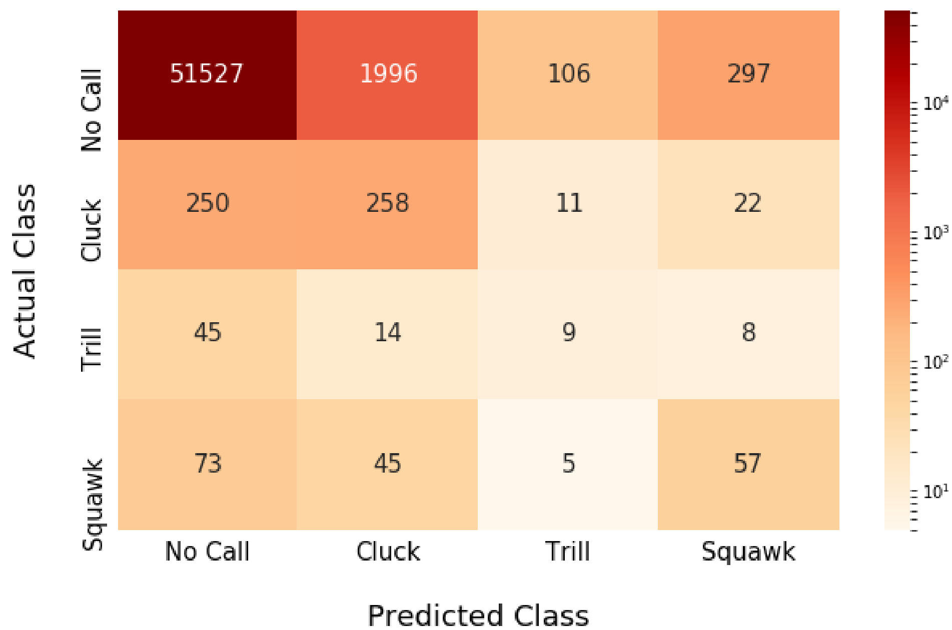
To reduce the number of false positives, I added deltas, and delta-deltas to my feature vectors. Deltas are estimates of the derivatives of each log-Mel feature, and are added to a model to increase temporal information for the isolated data points. Deltas are calculated using the following formula [4]:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

Where N is the number of neighboring features to either side of the sample for which we are calculating the delta. The code to calculate deltas was provided by the Company.

I chose a value of N=2, so to estimate the derivative of each log-Mel feature, the two samples preceding and following the feature were used. Calculating the deltas for each of the log-Mel features resulted in additional features being added to the data. I also added the deltas of the deltas (second derivative). The upsampled model with deltas showed a significant improvement over the upsampled model without deltas. The accuracy of the model increased from 85.6% to 94.8%, and the macro-weighted F1-score increased from 0.281 to 0.361.

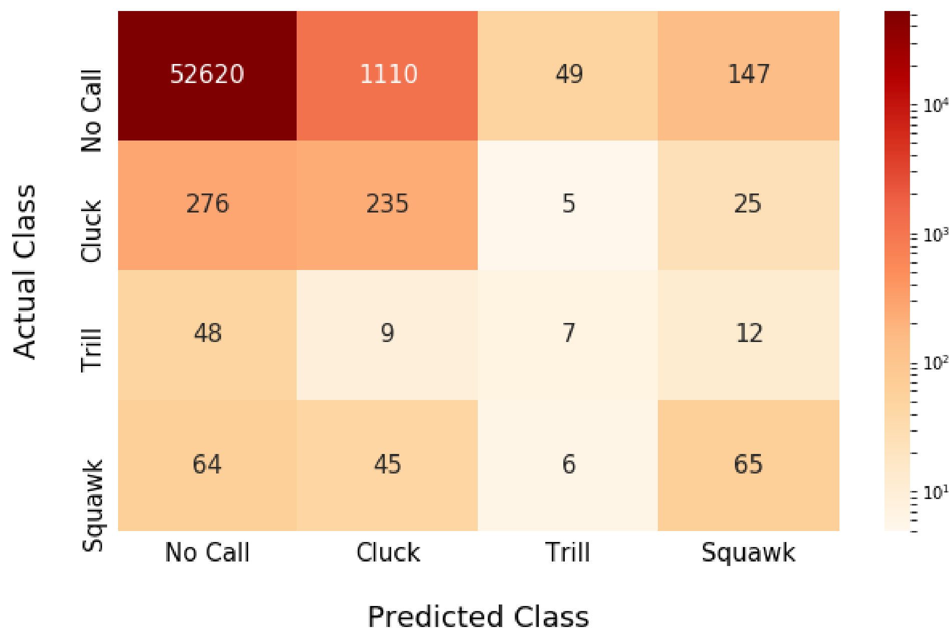
Table 3: Confusion matrix for model with deltas



Noise Reduction

To reduce the effect of background noise, I processed the audio signal with the noise reduction algorithm outlined in the Data section. I calculated log-Mel energy features and deltas from the noise-reduced signals, and upsampled the data to train a new model. The training and test data was also run through the noise reduction algorithm. Using noise-reduced data improved the performance of the model. Classification accuracy increased from 94.8% to 96.7%, and the macro-weighted F1-score increased from 0.361 to 0.407. The model trained on the noise-reduced data had slightly lower recall (0.466 vs. 0.467), but precision greatly improved (0.382 vs 0.330), indicating that noise reduction helped reduce the number of false positives.

Table 4: Confusion matrix for noise-reduced model



<u>Day/Night and Age Variables</u>

My teammate created variables to represent the age of the birds, and whether the recordings were taken during the day or at night. The age variable was introduced because bird vocalizations change as birds age, and an age indicator can allow XGBoost to split samples based on age. The day/night indicator was added because large fans run throughout most of the day to dissipate heat and moisture, which affects daytime audio quality. I included these variables in my final model.

<u>Final Classification Model</u>

My final classification model was trained on noise-reduced upsampled data. The features used were log-Mel energies, deltas, age, and a day/night indicator variable. The addition of the age and day/night variables dramatically improved the performance of my final classifier. My final model achieved a 97.7% classification accuracy, and a macro-weighted F1-score of 0.507. My model's macro-weighted F1-score is somewhat inflated by the 'no call' class, for which my model scores 0.989. The F1-scores for the vocalization classes are all much lower. The final model classification report is shown in Table 5, and the final model confusion matrix is shown in Table 6 below.

Table 5: Classification report for final model

	Precision	Recall	F1-Score
No Call	0.993	0.985	0.989
Cluck	0.254	0.464	0.328
Trill	0.304	0.276	0.290
Squawk	0.407	0.439	0.422
Macro Avg.	**0.490**	**0.541**	**0.507**

Table 6: Confusion matrix for final model

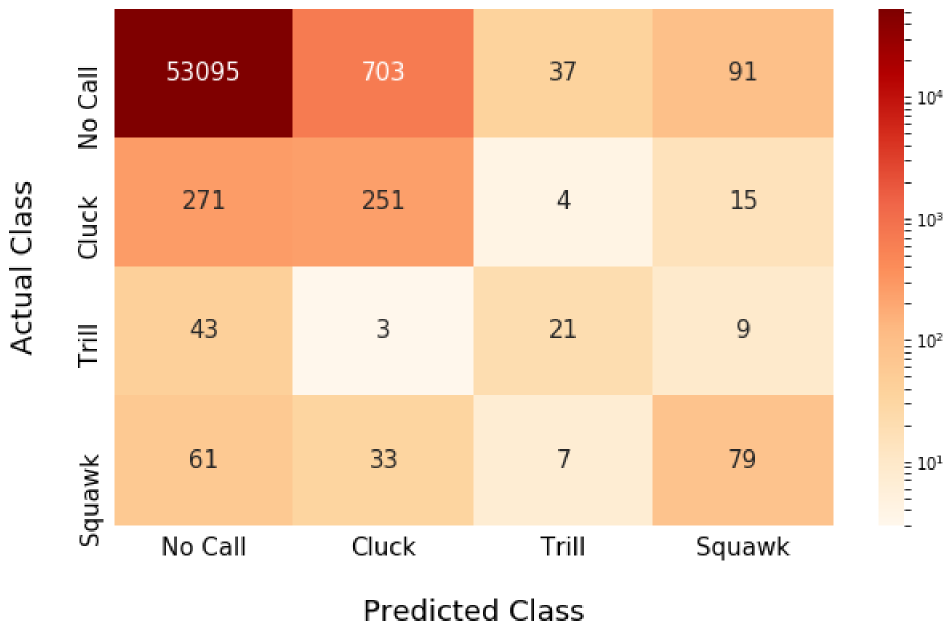


Table 7: Model metrics comparison

Metric	Basic Model	Upsampled Basic Model	Upsampled with Deltas	Upsampled, Deltas, Noise Reduction	Final Model
Accuracy	0.985	0.856	0.948	0.967	0.977
Precision	0.246	0.276	0.33	0.382	0.49
Recall	0.25	0.491	0.467	0.466	0.541
F1-Score	0.248	0.281	0.361	0.407	0.507

The final model demonstrates superior performance to each iteration before it, with the highest precision, recall, and F1-scores of any of the models, and only slightly worse accuracy than the basic model. I implemented upsampling so that my model would classify samples with labels other than ‘no call’, but upsampling introduced a large number of false positives and greatly reduced accuracy. The addition of deltas, noise reduction, age, and a day/night indicator was able to restore model accuracy, and classify more vocalizations correctly, while greatly reducing the number of false positives introduced through upsampling.

Classification Takeaways and Difficulties

While adding variables and tuning hyperparameters resulted in much better performance than the basic XGBoost model, my classifier still has significant room for improvement. There are a couple factors that may contribute to the difficulty our team had in correctly classifying chicken vocalizations:

1. There is persistent background noise from the large fans used to cool the chickens. This makes the manual labeling of data difficult, and can obstruct the underlying signal of vocalizations when engineering features.
2. Three individuals separately labeled files. While we spent a great deal of effort coordinating our labeling scheme, identifying sounds by ear is ultimately subjective. Inconsistencies in data labeling may render accurate classification difficult.

Mortality

***Please note that many of the dates and measurements have been removed for the purpose of confidentiality**

Predicting Audio Labels with XGBoost

I used my final XGBoost classifier to predict the labels for audio recordings of a separate flock from the training data. The audio data used for prediction was from the same exact microphone and same bird ages as the training data. Because my model was trained on noise-reduced data, I applied the noise reduction algorithm to each audio file before predicting labels. The predicted labels were summed by day and compared to daily mortality and controller (water consumption, feed consumption, temperature, etc.) data. A few of the audio files were corrupted and would not load correctly, so instead of raw counts, I converted the predicted counts for each class to percentages using the following equation:

$$percentage_{t,c} = \frac{count_{t,c}}{\sum_{c=1}^{C} count_{t,c}}$$

Where $count_{t,c}$ is the total number of samples classified as class=c for day=t.

The plot of vocalization percentages over time is shown in Figure 3. Immediately, it is clear that 'cluck percentage' may have some outliers. It is unlikely that the birds would increase their clucking frequency astronomically for a few days, before dropping back to near-previous levels. A possible explanation is that my classifier is classifying other sounds, such as feeding, drinking, or fan noise, as clucks for that time period. To test for the presence of outliers, I used the Hampel test, which identifies outliers as values that are significantly above and below a rolling median. The cluck counts for two dates were classified as outliers, because they were three standard deviations above the rolling median, so those counts were replaced with the rolling median values. The dotted purple line shows where the outliers were replaced. In general, clucks increase with age, while squawks and trills (the trend will be more obvious with the scaling of Figure 4) show a decreasing trend over time.

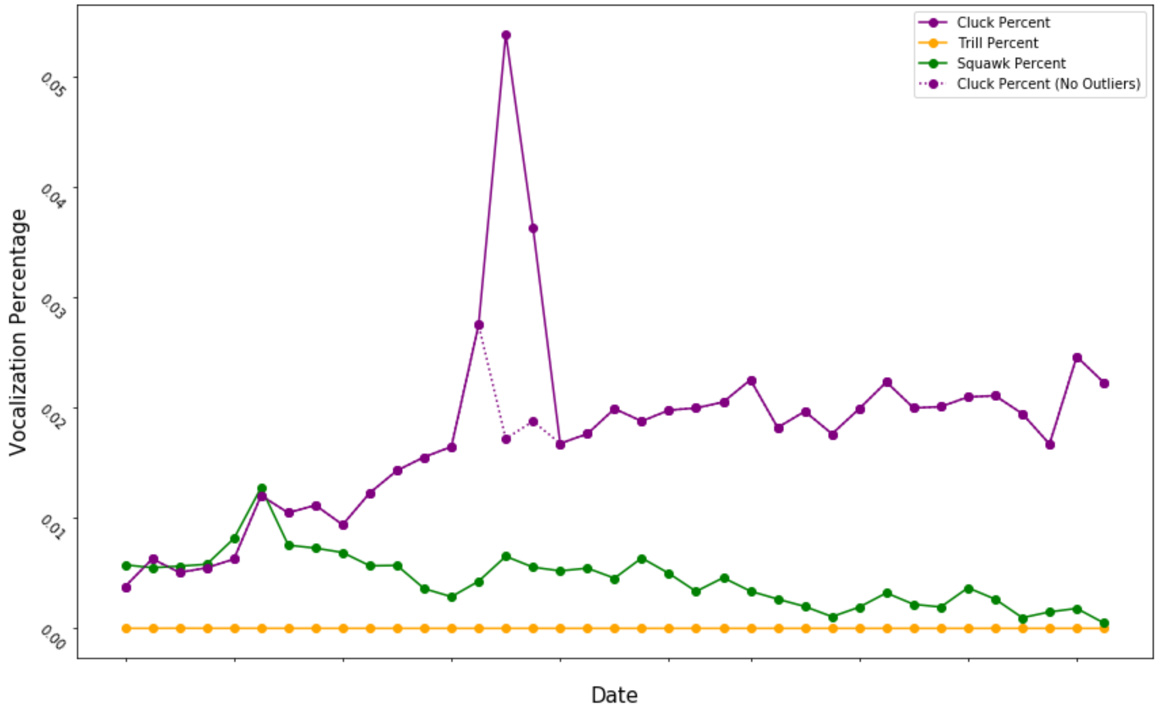


Figure 3 — Predicted vocalization percentages over time

Figure 4 shows plots of each of the time series variables in the data set.

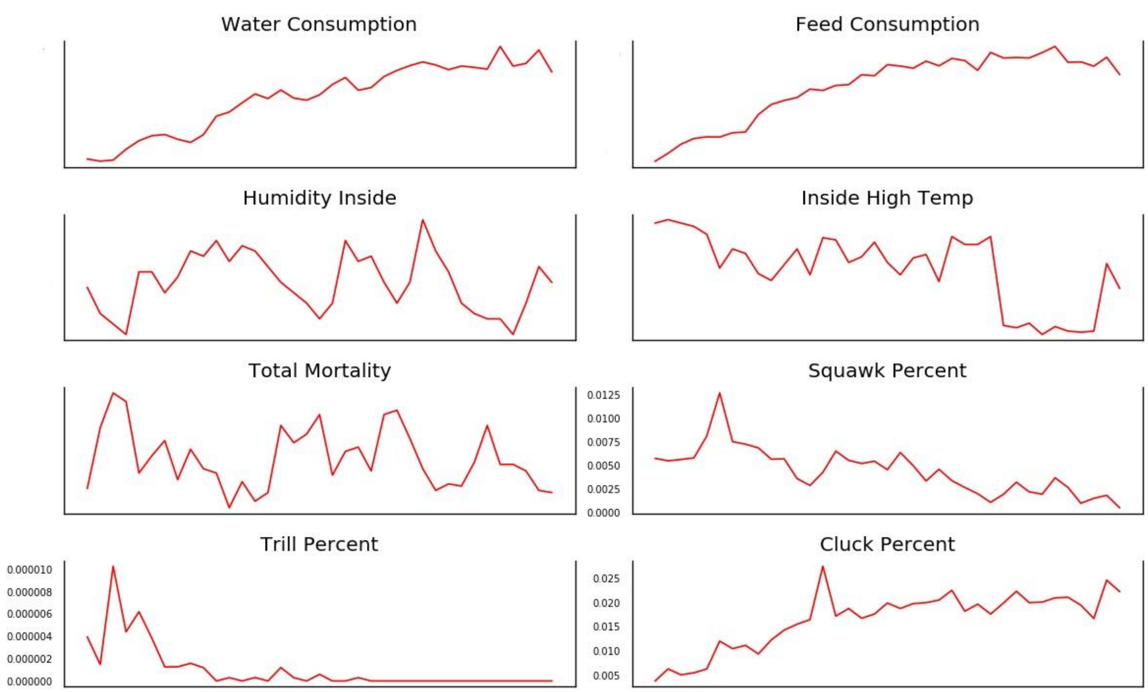


Figure 4 — Time series variable plots

My primary interest in vocalizations was exploring their relationship with mortality. Total mortality has a much higher variance than the vocalization percentages, and does not show a clear trend over time. Based on Figure 11, there does not appear to be an obvious relationship between total mortality and any of the predicted vocalization labels.

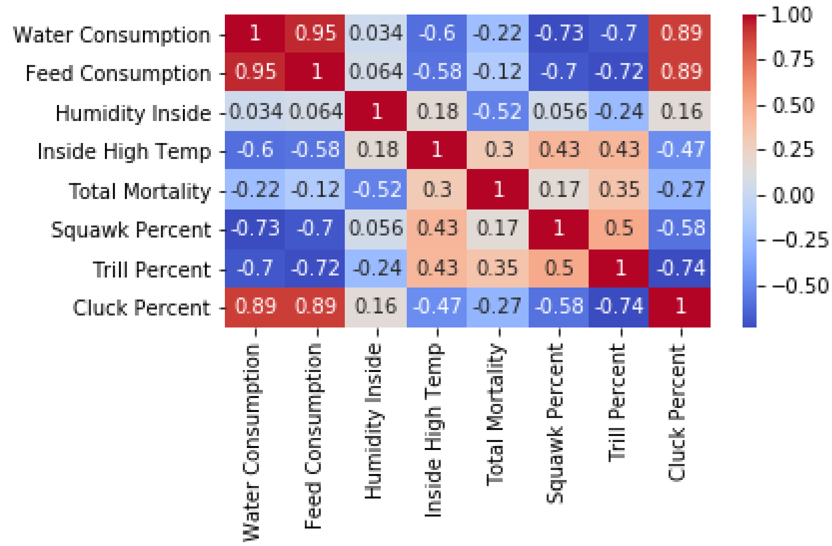


Figure 5 — Variable correlation coefficient heatmap

In Figure 6, I plotted cluck percent on the left axis and water consumption (in gallons) on the right axis. The two appear strongly correlated, as supported by a correlation coefficient of 0.89, shown in Figure 5. Cluck percent is also strongly correlated (0.89) with feed consumption.

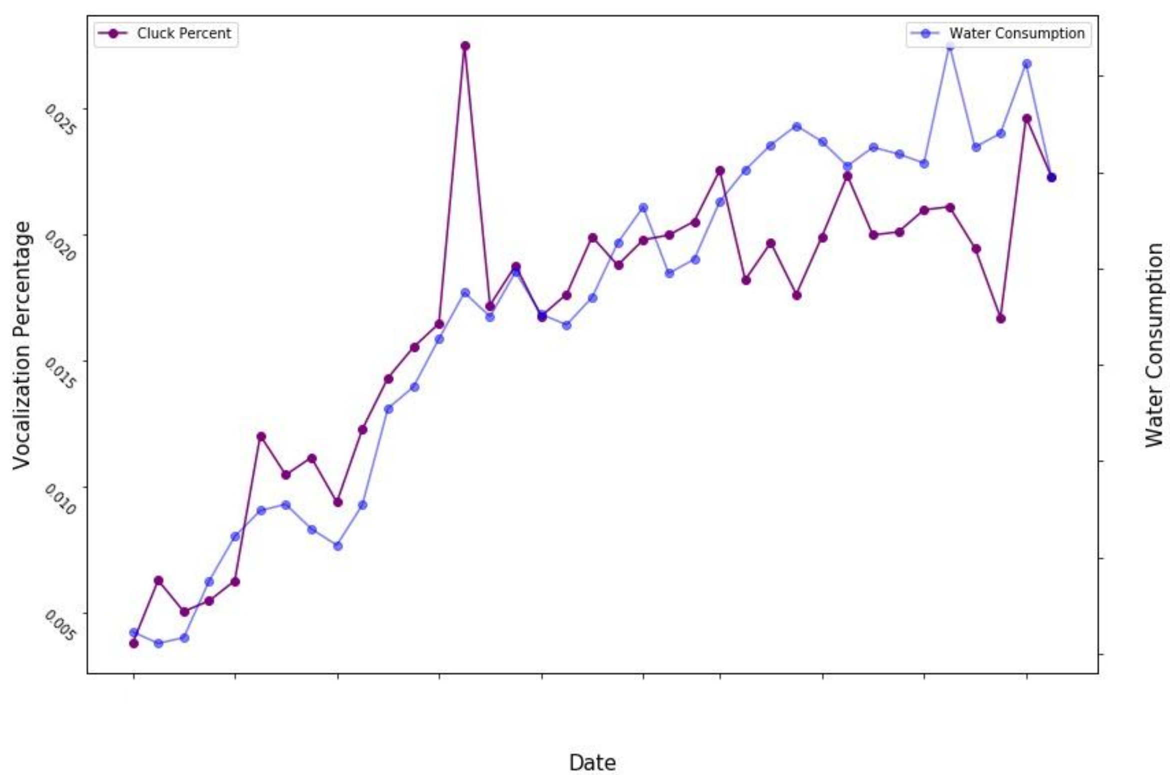


Figure 6 — Daily cluck percent and water consumption

To reduce multicollinearity, and because feeding and drinking behavior may be misclassified as clucks, I dropped feed consumption and cluck percent from my analysis. I kept water consumption over feed consumption because the Company has identified water consumption as a key variable in mortality analysis. The final variables were water consumption, humidity inside, inside high temperature, squawk percent, trill percent, and total mortality. The strong correlation between feed, water, and clucks could also be due to the fact that all three are indicators of general activity, but I do not feel that clucks add enough information that is not already included in water consumption to ignore the potential misclassification issues. Manual review of the cluck classifications is an area for future investigation.

Vector Autoregression Mortality Model

Before fitting the vector autoregression model, I ensured that all variables were stationary, which means that they have consistent properties (mean, variance, autocorrelation) over time [6]. I applied the Augmented Dickey-Fuller (ADF) test to each variable. The null hypothesis of the ADF test is that the data is non-stationary. For each variable, if the null hypothesis was not rejected, I differenced the variable (replaced the current day's value with the difference between the prior day's and current day's value), and applied the ADF test again. Water consumption, inside high temperature, and squawk percent required one round of differencing. Humidity inside, total mortality, and trill percent did not require any differencing.

After ensuring all variables were stationary through differencing, I fit a Vector Autoregressive (VAR) model on the time series data. The model chosen was VAR(lags=2), which means that the current value of each variable is a function of the previous two days of all variables.

The model output for total mortality is shown below in Table 26." L1.Squawk Percent" was significant at the p=0.05 level. "L1." refers to lag=1, or the previous day's value of the variable. Recall that squawk percent was differenced to achieve stationarity. This implies that an increase in squawk percent from two days ago to yesterday increases mortality today.

Table 8: Vector autoregression equation for total mortality

Results for equation Total Mortality

	coefficient	std. error	t-stat	prob
const	14.455072	48.153724	0.300	0.764
L1.Water Consumption	-0.035809	0.027509	-1.302	0.193
L1.Humidity Inside	-0.367944	0.536023	-0.686	0.492
L1.Inside High Temp	0.864381	1.066537	0.810	0.418
L1.Total Mortality	0.345678	0.213712	1.617	0.106
L1.Squawk Percent	2964.252531	1432.623542	2.069	0.039
L1.Trill Percent	1365039.757565	1258133.685593	1.085	0.278
L2.Water Consumption	-0.015820	0.026386	-0.600	0.549
L2.Humidity Inside	0.503601	0.543085	0.927	0.354
L2.Inside High Temp	1.552402	1.167202	1.330	0.184
L2.Total Mortality	0.151787	0.223815	0.678	0.498
L2.Squawk Percent	1888.546313	1609.265332	1.174	0.241
L2.Trill Percent	-1877250.379997	1278525.345262	-1.468	0.142

To check for significant effects on total mortality, I used the Granger Causality Test and impulse responses. The Granger Causality Test checks if values of a variable X provide statistically significant information about the response variable Y. None of the tested variables were able to reject the null hypothesis that they do not Granger-cause total mortality. The p-values for the Granger Causality tests are shown in Table 30 below.

Table 9: Granger Causality Test p-values

Granger Causality Test on Total Mortality

Variable	p-value
Water Consumption	0.422
Humidity Inside	0.636
Inside High Temp	0.375
Squawk Percent	0.090
Trill Percent	0.299
Total Mortality	0.126

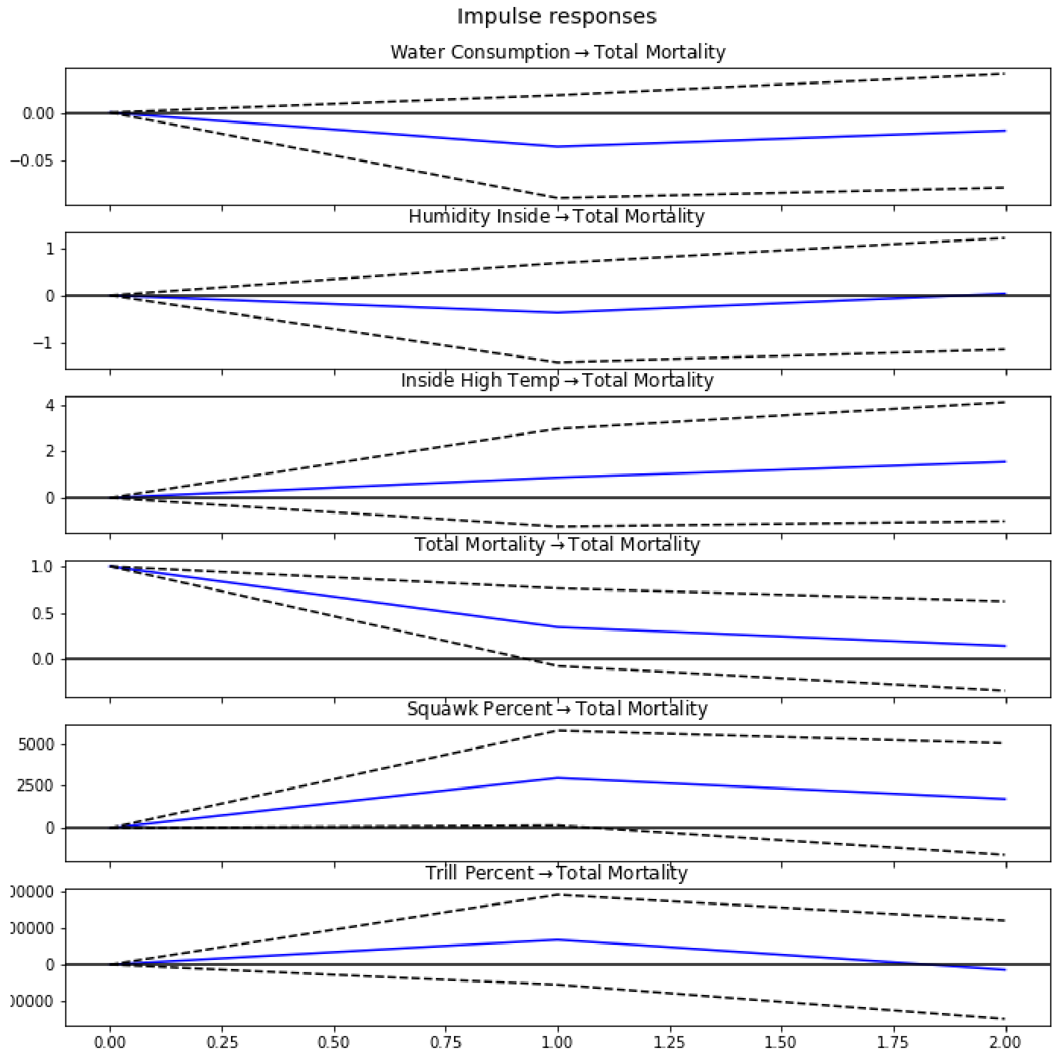


Figure 7 — Impulse response plots for total mortality

The impulse response plots in Figure 7 show the response of total mortality to a one unit shock in a given variable. The x-axis represents the number of days since the shock. The blue line represents the impact on total mortality, and the dotted lines above and below represent a 95% confidence interval. Other than total mortality, which is a given (a one unit increase in percent mortality will increase total mortality by one at t=0), only squawk percent shows a statistically significant effect on total mortality at any time. This can be shown by the fact that the lower confidence band crosses zero from t=0 until t=1 in the total mortality and squawk percent charts, but no other confidence bands cross zero.

The VAR model was able to find a weak association between squawk percent and total mortality. The regression coefficients and impulse responses show that L1.Squawk Percent is significant at the p=0.05 level. Squawks can indicate distress, so it follows intuition that an increase in squawking leads to increased mortality; however, the Granger causality test is not able to show a significant relationship at the p=0.05 level. The VAR model was built under the assumption that the classifications for squawk and trill were accurate. Based on the results reported for my classification model, this is likely a generous assumption, so these results must be considered under that context.

Sample size is the most challenging issue with flock mortality data. My audio classification model only labeled mid and late-stage samples, which brought the data down to 37 days. Differencing the variables to achieve stationarity removed another day. Fitting a VAR(2) model removed another two days due to the two-day lag required, so the final sample size was 34 days.

Another potential problem with the mortality data is that total mortality is created by the poultry farmer. Most of the mortalities recorded were chickens that were removed from the flock due to visible injury, deformity, or illness, and were not chickens found dead in the coop. Due to the subjective nature of the culling process, it is fair to wonder if the criteria for chicken removal is applied consistently each day.

Conclusions

Our team labeled 152 files of audio data with three different vocalizations and created multiclass classification models based on the labeled data. We also explored mortality data across two different flocks and created a simple model to predict mortality.

Our biggest classification challenge was dealing with imbalanced classes. We explored several methods to overcome this, including the creation of additional variables and upsampling our vocalization classes. These techniques helped create better models, but there is still significant room for improvement. Additional variables beyond the day/night indicator and age variables could help in this regard. Using other models that incorporate focal loss as a loss measure, such as a neural network, could improve classification accuracy. Using neural networks directly on Mel-spectrograms (instead of engineered features) has been implemented with success in bird vocalization classification competitions, such as BirdClef 2021 [1].

One recommendation to help improve the accuracy of audio classification models is to develop a universal vocalization taxonomy, along with numerous audio examples, that could be used as a basis for labeling efforts moving forward. Our team spent considerable time developing our labeling scheme, and there may have been inconsistencies in our labels. Having a framework to work off of could increase time spent modeling, and improve the consistency of the audio labels. It would also allow labels to be aggregated across several teams, increasing the size of the data set, which would also improve classification model accuracy.

We built a multivariate time series model to predict mortality using labels generated by our XGBoost model. We found some evidence that squawks, which may indicate distress, are predictive of future mortality.

The multivariate time series regression approach can be applied to other flocks to see if the results hold with other data. Another possible approach to modeling mortality is the use of a Long Short-Term Memory (LSTM) Neural Network. Determining the effect that factors have on mortality is important to farmers, as it could encourage them to adjust the environmental conditions the chickens live under, or in the case of vocalizations, alert them that something is wrong. Future study in this area could provide significant benefits to farmers and the Company.

References

1. *BirdCLEF 2021 - Birdcall Identification*. (2021, April 1). Kaggle. https://www.kaggle.com/c/birdclef-2021

2. *GridSearchCV.* Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

3. *KNN — Amazon SageMaker Examples*. (n.d.). Amazon Sagemaker. https://sagemaker-examples.readthedocs.io/en/latest/introduction_to_amazon_algorithms/k_nearest_neighbors_covtype/k_nearest_neighbors_covtype.html

4. *LightGBM.* https://lightgbm.readthedocs.io/en/latest/

5. Maka, T. (2015). Change Point Determination in Audio Data Using Auditory Features. *International Journal of Electronics and Telecommunications*, *61*(2), 185–190. https://doi.org/10.1515/eletel-2015-0024

6. Nau, R. (n.d.). *Stationarity and differencing of time series data*. Duke University. https://people.duke.edu/%7Ernau/411diff.htm

7. *Mel Frequency Cepstral Coefficient (MFCC) tutorial.* (n.d.). Practical Cryptography. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

8. *precision_recall_fscore_support*. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

9. *RandomForestClassifier*. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

10. *RandomizedSearchCV*. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

11. Roberts, L. (2020, March 14). *Understanding the Mel Spectrogram - Analytics Vidhya*. Medium. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

12. Shepon, A., Eshel, G., Noor, E., & Milo, R. (2016). Energy and protein feed-to-food conversion efficiencies in the US and potential food security gains from dietary changes. *Environmental Research Letters*, *11*(10), 105002. https://doi.org/10.1088/1748-9326/11/10/105002

13. Tonsor, G. T., & Olynk, N. J. (2011). Impacts of Animal Well-Being and Welfare Media on Meat Demand. *Journal of Agricultural Economics*, *62*(1), 59–72. https://doi.org/10.1111/j.1477-9552.2010.0026

14. XGBoost Algorithm. Amazon Sagemaker. https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html