Evaluation of Best Friends Animal Society's Predictive Models for No-Kill Shelters in the United States

February 7, 2025

Dr. Adam Feltz, Mx. Uyen Hoang, and Mx. Jenna Holt
The University of Oklahoma

Executive Summary: We conducted an independent review in February of 2025 of materials sent to us by Best Friends Animal Society (BFAS). We had 3 self-identified objectives in this evaluation: (1) replicate the findings of BFAS; (2) provide additional assessment of the models and results; (3) evaluate how well the models avoid false positives. The models, data, and code BFAS provided did very well on each of the three evaluative objectives. Overall, the approach that the BFAS takes is likely to be robust and provide accurate predictions about no-kill shelters in the United States.

1. Replication of regression models

In Section 1, we report our results about replicating the regression models using the data and code provided by the BFAS. We also provide some additional analyses involving 95% confidence intervals of the BFAS estimates along with our calculations for percent of accurately identified shelters that were no-kill or not no-kill shelters. These additional data will be useful in further evaluating the models in Sections 2 and 3.

Our first set of analyses was to recreate the correlations between the training set (80% of the data) and the test set (20% of the data). BFAS provided us with their code and their data. We replicated the results without issue and nearly exactly to the decimal point. We also took this opportunity to calculate the 95% confidence intervals for the correlations between the training and test set. The 95% confidence intervals will become useful to help interpret the results provided in Section 2. The correlations we calculated, along with their 95% confidence intervals, are in Table 1.

Table 1. Our calculations of correlations between training and test data along with 95% confidence intervals of those correlations.

Pre-pandemic	**Correlation between training and test set**	**95% Confidence Interval**
Gross Intake	0.89	.86, .92

Net Intake	0.84	.80, .88
Non-Live Outcomes	0.86	.83, .90
Post-pandemic		
Gross Intake	0.93	.91, .95
Net Intake	0.93	.91, .95
Non-Live Outcomes	0.91	.87, .93

We also calculated the number of shelters that were accurately identified as no-kill and shelters that were accurately identified as not no-kill shelters (see Table 2). The total number of correctly identified shelters is nearly identical to those reported by the BFAS again suggesting that we could re-create their results. The specific number of correctly identified no-kill and kill shelters, rather than total number of correctly identified shelters, will become important to the analyses in Section 3.

Table 2. Our estimates of number of correctly identified not no-kill and no-kill shelters.

	Correctly Identified not No-Kill	Correctly Identified No-Kill	Our Estimate Total Correctly Identified	BFAS Total Correctly Identified
Pre-Pandemic	73	93	166	164
Post-Pandemic	90	110	200	198

2. Repeated random train-test set analyses

One worry with the approach BFAS took is perhaps there is something idiosyncratic about the training set of data (80% of the data) or the test set data (20% of the data) that could bias or otherwise introduce error into the estimates. One common way to help address that worry is by repeating the analyses ("trials") but randomly assigning data to the training and test set for each trail. This can provide an overall estimate of the outputs of a model that do not depend on one specific assignment of data to the training or test sets and can give a sense for the performance of the model on each trial.

We used 10 iterations of randomly assigning 80% of the data as the training set and 20% to the test set. We then calculated average correlations for the models used by the BFAS (see Table 3). Responses were in line with what we would expect given the original analyses and

results. The average correlations of our analyses results fall within the 95% confidence interval of the originally reported correlations (see section1). These results again suggest that the model, data, code, and approach employed by the BFAS is likely to be robust and accurate.

Table 3. Average correlation of the models of key outcome variables using based on 10 iterations.

Pre-pandemic	**Average correlation between training and test sets**
Gross Intake	0.89
Net Intake	0.86
Non-Live Outcomes	0.84
Post-pandemic	
Gross Intake	0.94
Net Intake	0.93
Non-Live Outcomes	0.93

3. Precision, Recall, and F1 Scores

Since one BFAS's stated goals was not only average accuracy but also not incorrectly identifying a not no-kill shelter as no-kill (i.e., avoiding false positives), we conducted one final supplemental analysis that involved calculating recall, precision, and F1 scores. Recall scores give a sense for how well a model can identify positive results correctly without paying attention to potential false positives (i.e., maximizes finding true positives). Precision refers to how well a model can identify positives results that are actually positive even if some potential true positive results are missed (i.e., minimizes reporting false positives). F1 scores combines those two measures to give a sense for how well a model can identify positives without incorrectly identifying negatives.

The BFAS stated that they prefer conservative estimates of their models and want to minimize false positives even at the expense of perhaps missing some true positives. For the purposes of the BFAS, the precision score is perhaps the most relevant. We provide all three scores for each trial that we conducted for the analyses in Section 2 in Table 4. In these analyses, we used a fixed rate of 45% of the correctly identified shelters being no-kill and 55% of the correctly identified shelters being kill shelters for all trials. These percentages were based on our

estimates of accurately identified not no-kill and no-kill shelters in Section 1. The mean recall score was .81 (i.e., 81% of no kill shelters were identified), the mean precision score was .86 (i.e., 86% of the shelters identified as no kill were in fact no kill), and the mean F1 score was .83. All the values would be considered strong in many contexts (e.g., a value greater than .8).

Table 4. results for the repeated randomized trial analyses along with the recall, precision, and F1 scores for each trial.

Trial	Period	False Negative	False Positive	Match	False Negative (%)	False Positive (%)	Match (%)	Recall	Precision	F1
1	Pre-pandemic	15	20	164	8	10	83	0.83	0.78	0.80
1	Post-pandemic	19	21	188	9	10	85	0.81	0.80	0.81
2	Pre-pandemic	21	8	170	11	4	85	0.78	0.90	0.84
2	Post-pandemic	17	18	193	8	8	86	0.83	0.83	0.83
3	Pre-pandemic	18	17	164	10	9	83	0.80	0.81	0.80
3	Post-pandemic	17	19	192	8	9	85	0.83	0.82	0.82
4	Pre-pandemic	27	15	157	15	9	79	0.72	0.82	0.77
4	Post-pandemic	24	15	189	11	7	82	0.78	0.85	0.81
5	Pre-pandemic	21	12	166	11	6	82	0.78	0.86	0.82
5	Post-pandemic	16	12	200	7	6	87	0.85	0.88	0.86
6	Pre-pandemic	15	14	170	8	7	85	0.83	0.84	0.84
6	Post-pandemic	11	9	208	5	4	91	0.89	0.91	0.90
7	Pre-pandemic	21	10	168	11	5	84	0.78	0.88	0.83
7	Post-pandemic	15	14	199	7	7	88	0.85	0.86	0.86

8	Pre-pandemic	18	11	170	10	6	85	0.81	0.87	0.84
8	Post-pandemic	19	12	197	9	6	86	0.82	0.88	0.85
9	Pre-pandemic	19	15	165	10	8	82	0.79	0.83	0.81
9	Post-pandemic	15	10	203	7	5	89	0.86	0.90	0.88
10	Pre-pandemic	19	9	171	10	5	85	0.80	0.89	0.84
10	Post-pandemic	19	9	201	10	5	85	0.82	0.91	0.86

4. General Assessment and Recommendations

Our independent, general assessment suggests that the methods, conceptualizations, implementation, and current models are likely to be robust predictors of no-kill shelters while also minimizing false positives. Our analyses might provide some minor points to supplement their analyses. For example, the BFAS may also want to consider explaining in more detail how predictors in their models were selected (e.g., numerical R^2 increase or significant R^2 increase). We also encourage the BFAS to make good on their proposed future research plans including time-series and potential other predictors (although we are somewhat skeptical of better model fits since the models are already strongly predictive), or a smaller set of better predictors for simplicity and computational efficiency. Overall, our general assessment is that the BFAS is performing rigorous, scientifically and statistically responsible research predicting not no-kill and no-kill shelters.