**Regression Modeling of Adoption Returns**

**Introduction**

Logistic regression modeling was applied to the analysis of adoption returns in order to achieve two goals:

1. Understand factors that lead to higher risk of return
2. Generate a scoring system that could be used to identify adoptions at greater risk

Logistic regression has two advantages over more modern machine learning algorithms. It estimates probability rather than trying to classify outcome, and it returns an accessible model which can be adapted into an Excel spreadsheet.

**Methods**

The R package *‘rms*’was used for modeling. This package is created and maintained by Frank Harrell of Vanderbilt University, who is the author of *Regression Modeling Strategies*. The package is ‘one-stop shopping’ for logistic regressions allowing for curvature in predictors (for example, younger and older animals both being at increased risk) and for assessing whether the model is overfit (too many parameters) or has other limitations.

**Data Cleaning**

The ‘PetPoint\_byAnimal.csv’ file was restricted to Intake.Dates on or after 2017, so that extreme stays were excluded.

The ‘Intake.Condition’ field was recoded into two new flags, ‘Sick’ if Intake.Condition included the word “Sick”, and ‘Injured’ if Intake.Condition included the word ‘Injured’ – so ‘Sick and Injured’ would show up in both.

The Dates (Intake.Date, Outcome.Date and Date.of.Birth) were recoded into a consistent format. Animal gender, which included 16 unknowns, was recoded into a binary Female flag. Month of adoption was extracted from Outcome.Date.

Visits were sorted by date by animal. If a given visit had an Intake.Subtype that included “Return,” then the visit before it was flagged as Return.From. the adoption from which animals were returned.

In the ‘PetPoint\_byPerson.csv’ file, an outside\_philly flag indicated whether the ‘City’ field was not equal to ‘Philadelphia.’

The Animal and Person tables were linked on Animal.ID and on Outcome.Date. The final table had 6,628 fields. Surprisingly, there were 499 (7.5%) cases where the intake was a Return, but only 321 (4.8%) where there was a prior visit to be returned from. This was not a consequence of the Intake.Date limitation, but may reflect either recording issues or the animal not being returned to the same site it was adopted from.

**Model creation and pruning**

The initial model used Species (cat vs Dog), animal gender, person gender, sick, injured, outside\_philly, month of adoption, LOS and Outcome.Age.in.Months as predictors. Person gender and animal gender were allowed to modify the effect of other predictors. Splines were used to allow LOS and Outcome.Age to follow trends other than monotonic linear.

The ‘calibrate’ function in the ‘rms’ package assessed how well observed probabilities followed predictions while adjusting for model optimism, and the ‘validate’ functions used backward stepwise regression to prune the model. Both functions used bootstrap resampling (200 resamples) to look at model performance as the distribution of data was varied.

**Results**

The original model showed significant signs of overfitting, where bias-correction led to variation from the ideal line where observed probability met predicted probability (see Figure below)



The final model showed improvement especially at low risk of adoption returns



The final model dropped non-linearities and most prediction terms. The resulting model was:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | Estimated log-odds | Std error | Wald Statistic | p.value |
| (Intercept) | -3.17 | 0.13 | -24.3 | 1.09E-130 |
| Species.aDog | 1.02 | 0.222 | 4.6 | 4.26e- 6 |
| Gender.pF | 0.408 | 0.125 | 3.26 | 1.13e- 3 |
| Gender.pM | 0.216 | 0.177 | 1.22 | 2.22e- 1 |
| SickTRUE | 0.344 | 0.135 | 2.55 | 1.07e- 2 |
| LOS | -0.00542 | 0.00119 | -4.56 | 5.06e- 6 |
| Outcome.Age.in.Months | 0.007 | 0.00231 | 3.04 | 2.39e- 3 |
| Species.aDog:Outcome.Age.in.Months | -0.0134 | 0.00424 | -3.15 | 1.62e- 3 |

Generally, these are not modifiable. It is surprising that adopters with defined gender (F or M) seem to have higher risk of returns than adopters with undefined gender (0).

It may be informative that the risk of adoption returns decreases with LOS.



join\_tbl$Returned.From: FALSE

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 19.00 51.00 71.05 95.00 829.00

-----------------------------------------------------------------------------------------------------------------

join\_tbl$Returned.From: TRUE

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0 13.0 32.0 47.6 58.0 368.0

This suggests that adoption may be more successful if the animal is not adopted shortly after intake.

Also, the effect of animal age may be different from cats and dogs. Older cats and younger dogs are more likely to be returned. (The shaded region is 95% confidence interval on estimate).



FALSE

: Cat

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

1.00 4.00 8.00 19.89 26.00 186.00 55

---------------------------------------------------------------------------------

: TRUE

: Cat

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

1.00 4.00 13.00 24.64 37.00 114.00 1

---------------------------------------------------------------------------------

: FALSE

: Dog

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 16.50 41.00 52.17 82.50 171.00

---------------------------------------------------------------------------------

: TRUE

: Dog

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 12.00 29.00 42.03 60.50 145.00

**Conclusions**

The model does not seem sufficient to predict those adoptions which are likelier than not (i.e., > 50% probability) to end in a return. However, it may be useful to rule out those persons at particularly low risk for return and ensure adoptions in higher risk situations (shorter LOS, younger dogs, older cats) get more attention during the adoption process.