PetFinder Adoption Prediction

George Cruz Karim Hammoud Maliat Islam Gabriella Martinez Ken Popkin

Date: 2021-12-12 Due: 2021-12-12

Contents

1 Overview

There are millions of stray pets around the world, some of which are fortunate enough to be adopted while many others are not. While adoption of a pet is often the definition of success, the rate at which a pet is adopted is also a key success factor - pets that take a long time to adopt contribute to over-crowded animal shelters and can prevent taking on new strays. Sadly, pets that are not adopted eventually need to be euthanized.

1.1 Learn more about the data

[1](https://www.kaggle.com/c/petfinder-adoption-prediction/data)

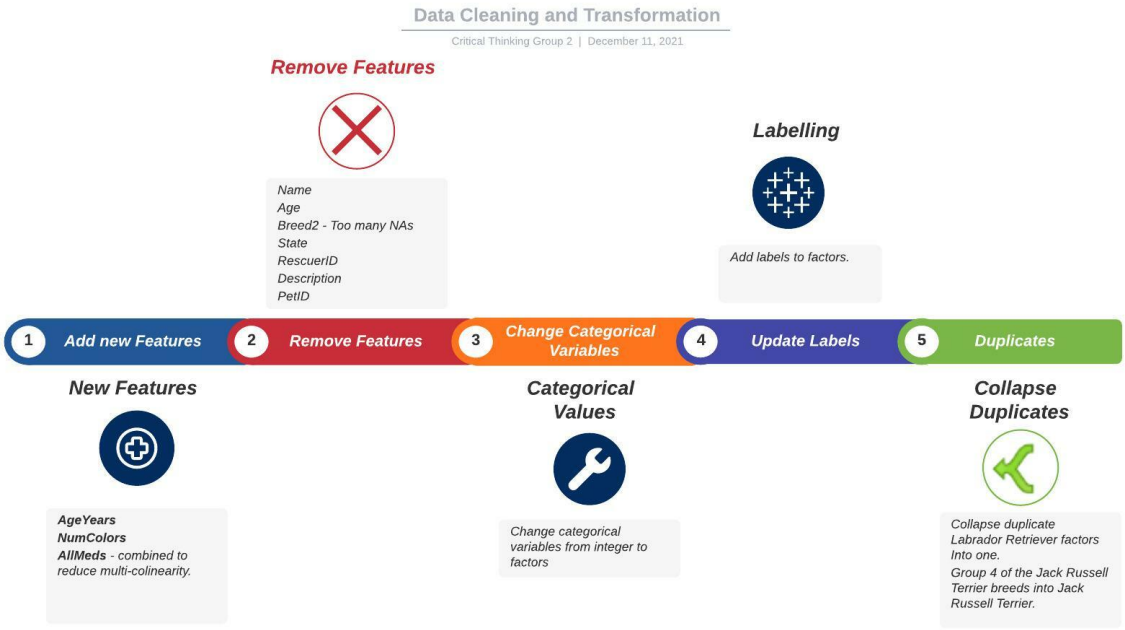
2 What to Predict

Predictor (Adoption Speed) Description: Predict how quickly, if at all, a pet is adopted.

The values are determined in the following way: 0 - Pet was adopted on the same day as it was listed. 1 - Pet was adopted between 1 and 7 days (1st week) after being listed. 2 - Pet was adopted between 8 and 30 days (1st month) after being listed. 3 - Pet was adopted between 31 and 90 days (2nd & 3rd month) after being listed. 4 - No adoption after 100 days of being listed.

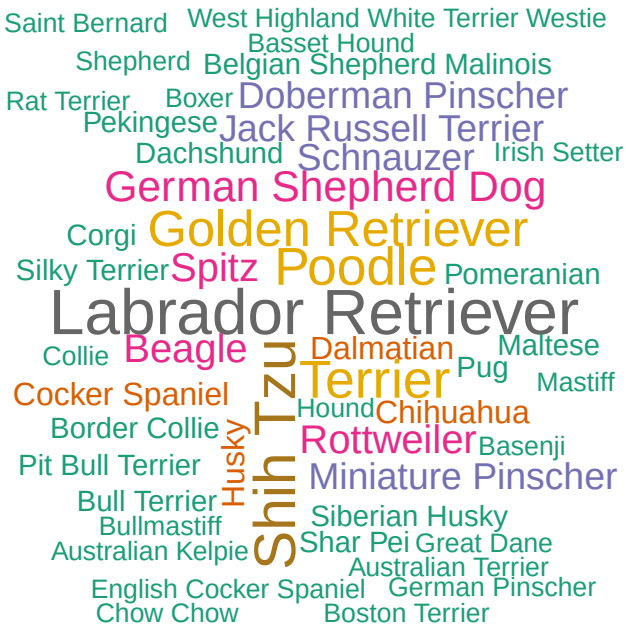
[1]About the data

3 Data Cleaning and Transformation



4 Dog Breed Word Cloud

Below we use `wordcloud` to visualize the dog breeds in our data. Note the Mixed Breed has been removed as it was skewing our data with the highest number of observations with 5923 total.



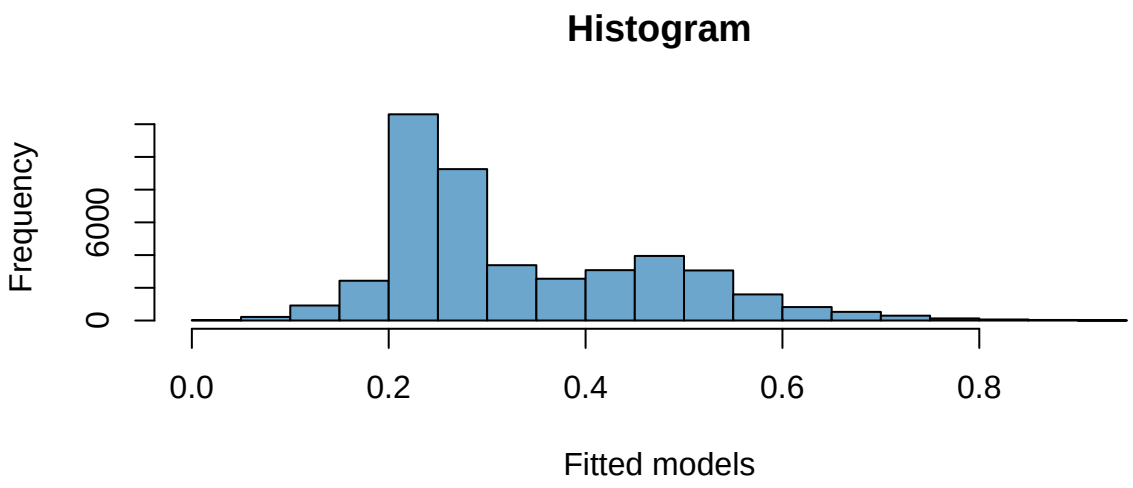
5 Ordinary Logistic Regression Model

We created several models using ordinal logistic regression since response variable can be considered an ordinal factor. Ordinary Logistic Regression

```
## Call:
## polr(formula = euth_risk ~ Type + (Gender + Color1 + MaturitySize +
##     FurLength + Health + Quantity + Fee + PhotoAmt + AgeYears +
##     NumColors + purebreed + AllMeds), data = train_data2, Hess = TRUE)
##
## Coefficients:
##                            Value Std. Error t value
## TypeDog                 0.6465752  0.0539215 11.9910
## GenderMale              0.0372710  0.0344366  1.0823
## GenderMixed             0.0082918  0.0618420  0.1341
## Color1Brown            -0.0076983  0.0402321 -0.1913
## Color1Cream             0.1009893  0.0696028  1.4509
## Color1Golden            0.0965965  0.0676829  1.4272
## Color1Gray              0.0995334  0.0797473  1.2481
## Color1White             0.1525216  0.0835323  1.8259
## Color1Yellow           -0.0930997  0.0796500 -1.1689
## MaturitySizeLarge      -0.9301085  0.3582734 -2.5961
## MaturitySizeMedium     -0.7983518  0.3556510 -2.2448
## MaturitySizeSmall      -0.8899883  0.3564297 -2.4970
## FurLengthMedium        -0.2220177  0.0715048 -3.1049
## FurLengthShort         -0.2108798  0.0706601 -2.9844
## HealthMinor Injury     -0.1949782  0.0892682 -2.1842
## HealthSerious Injury   -0.0990709  0.3472700 -0.2853
## Quantity               -0.0880270  0.0153361 -5.7399
## Fee                    -0.0006377  0.0002148 -2.9696
## PhotoAmt                0.0804304  0.0052383 15.3543
## AgeYears               -0.1227725  0.0124293 -9.8776
## NumColors2              0.1539906  0.0410001  3.7559
## NumColors3              0.1028811  0.0486792  2.1135
## purebreedYes            0.6861984  0.0554029 12.3856
## AllMeds4                0.5020650  0.0551127  9.1098
## AllMeds5                0.3858815  0.0564949  6.8304
## AllMeds6                0.4401485  0.0537605  8.1872
## AllMeds7                0.1436685  0.0948477  1.5147
## AllMeds8                0.7181419  0.1009836  7.1115
## AllMeds9               -0.1911094  0.0732132 -2.6103
##
## Intercepts:
##              Value   Std. Error t value
## High|Low    -0.7194  0.3711    -1.9386
## Low|Medium   0.3275  0.3711     0.8827
##
## Residual Deviance: 30510.98
## AIC: 30572.98
## (5 observations deleted due to missingness)
```

6 OLR Histogram

After dealing with multicollinearity and insignificant values, we see that the model with the lowest AIC value of the models created is `model_2` with an AIC of 29952.42.



7 OLR Predictions

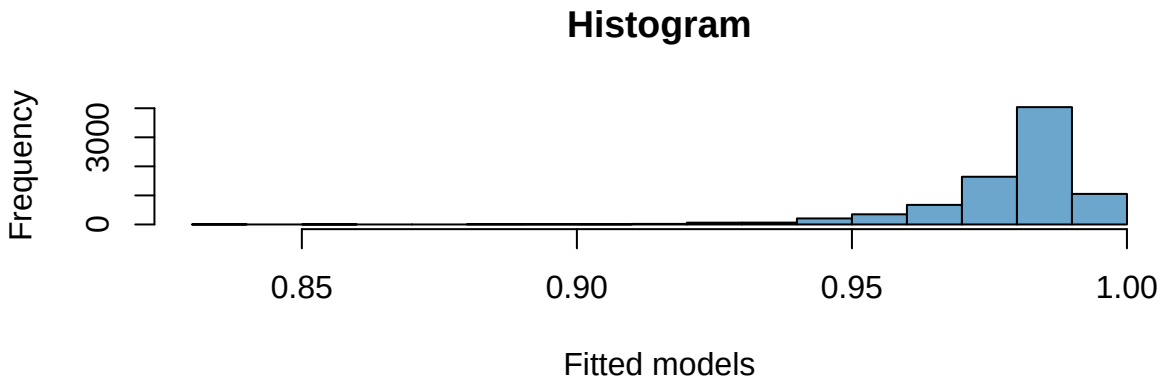
Below we generate our predictions with the `predict` function Toward Data Science - OLR.

```
##        High       Low    Medium
## 1 0.2930491 0.2484305 0.4585204
```

This model predicts that our Evaluation_Pet will be a medium risk of euthanasia where it will adopted in either months 1, 2 or 3 of being listed.

8 Binomial Logistic Regression Model

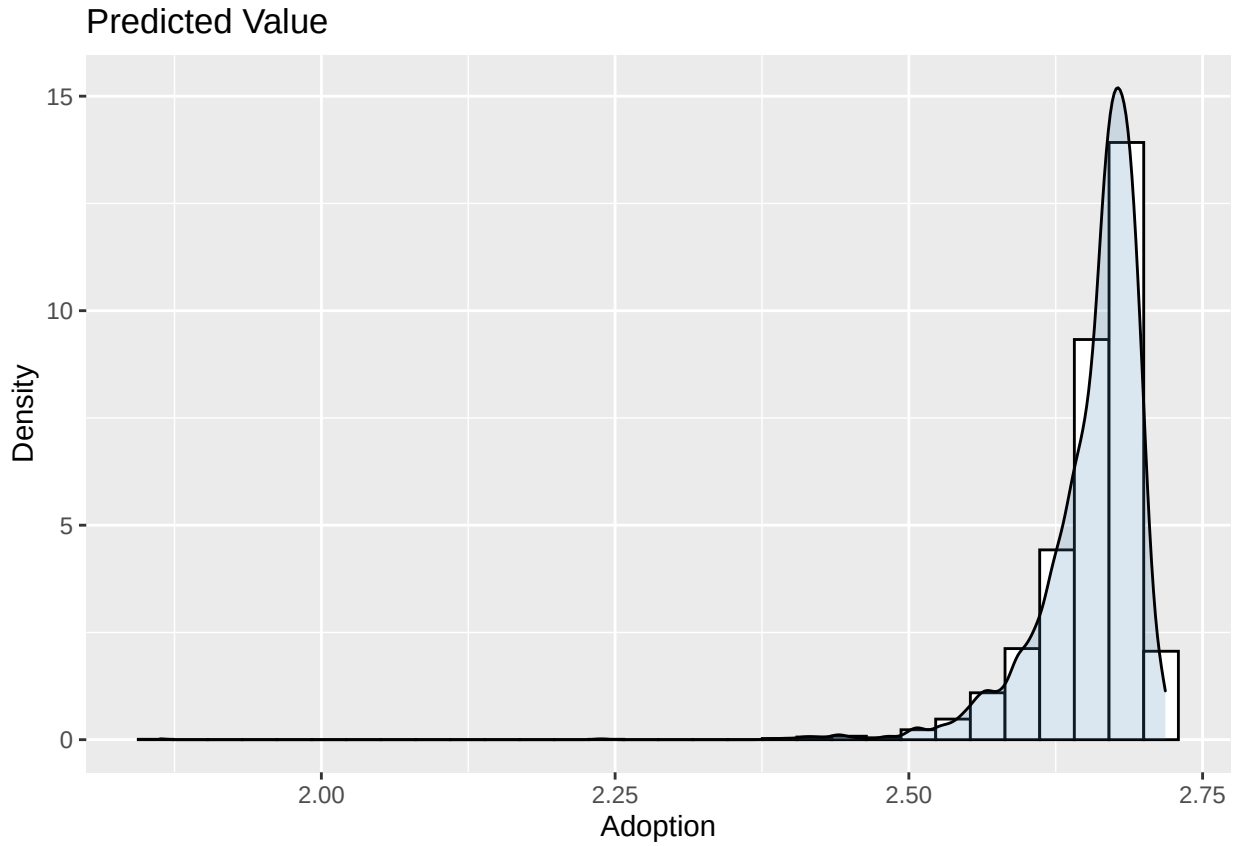
We built a Binomial Logistic Regression Model using a subset of the training data of only dogs. After using the Step Method we arrived at the best fitted model.



Here we can see the model works really well on the train data, we need to use it on the evaluation data and check the accuracy of the model on it.

9 BLR Predictions

When we apply the model to the evaluation data we get the following results.

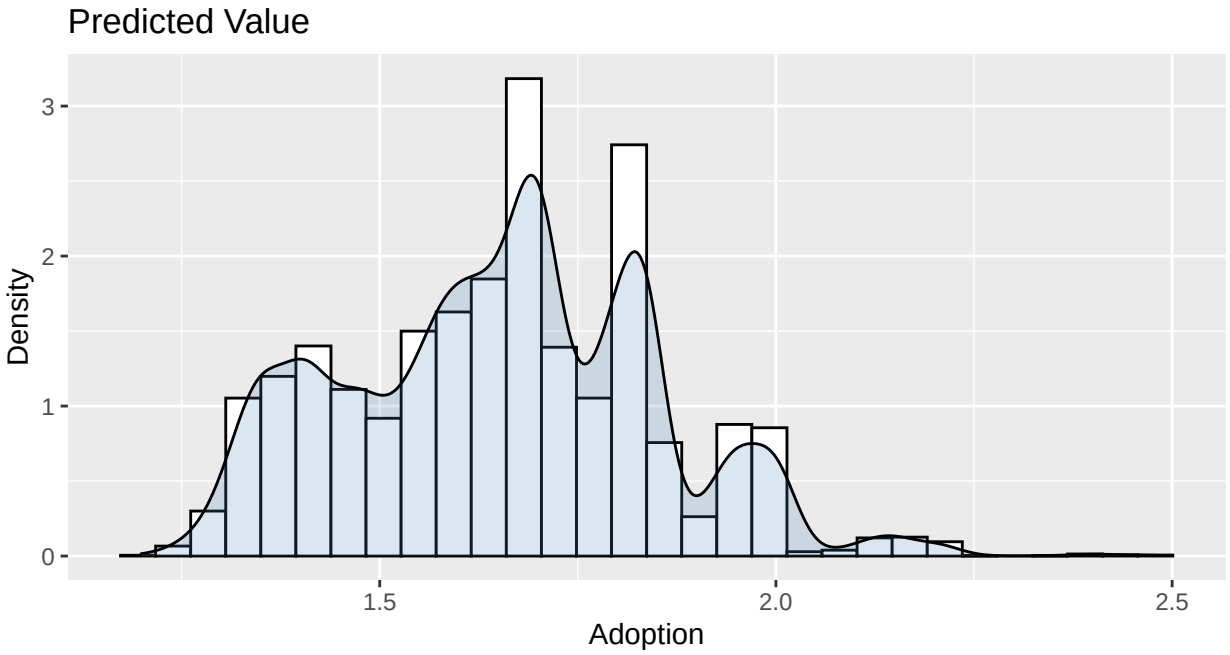


We can see that the adoption values is more skewed to the right, the density of the adoption rate is more around 2.5 to 2.75

10 Negative Binomial Model

We also trained several Negative Binomial Models and found the best fit would be the one that uses a few significant variables.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1813  0.4076  0.5103  0.4952  0.5852  0.9092
```

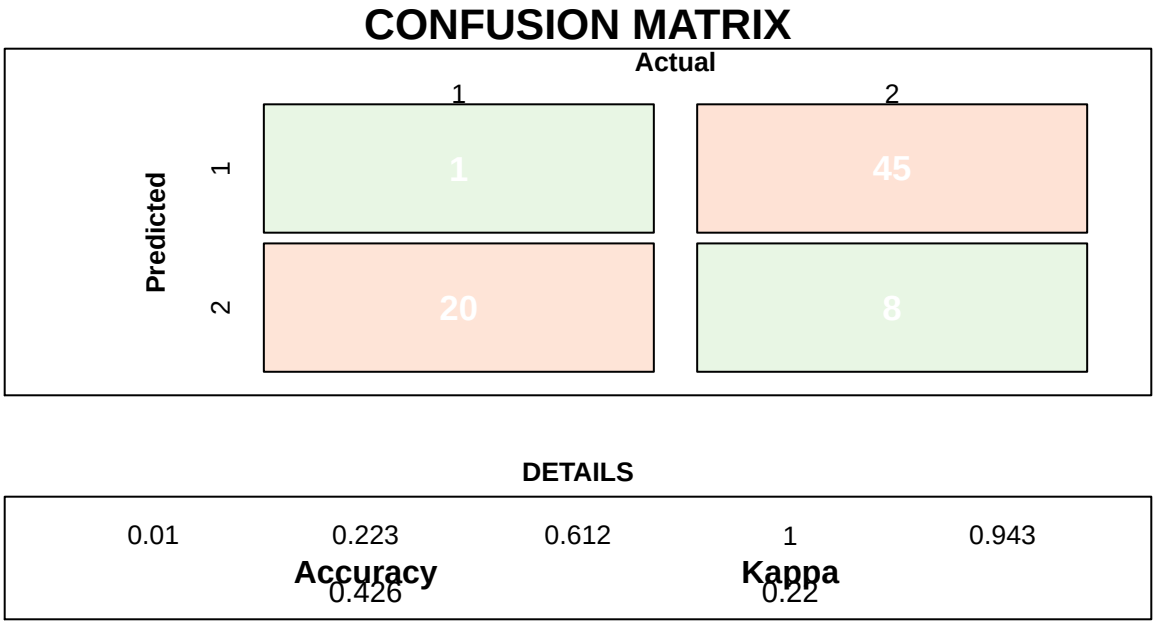


It can be observed that based on the prediction that Mean is 0.4953 and Median is 0.518.

11 Random Forest, XGBoost

Helpful article from TowardsDataScience

We explored other models like Random Forest and XGBoost. However, we were not able to get a better fit than with Binomial or OLR. Here a confusion matrix from one of our RF Models:



12 Conclusions

We tried several models: - Ordinary Logistic Regression - Binomial Logistic Regression - Negative Binomial - Linear Modeling - Random Forest - XGBoost

With the data provided, our best fitted models were the **OLR** and the **BLR**.

13 Links and References

13.0.1 References

Kaggle - About the data. "https://www.kaggle.com/c/petfinder-adoption-prediction/data" UCLA - Ordinary Logistic Regression. "https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/" Towards-DataScience - Random Forest in R. "https://towardsdatascience.com/random-forest-in-r-f66adf80ec9" Toward Data Science - OLR. "https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5".

13.0.2 Notable Links

Project Github. "https://github.com/akarimhammoud/Data_621/tree/main/Final%20Project" Visualizations. "https://rpubs.com/krpopkin/846476"