Predicting Animal Adoptability from Key Indicators

Team Members:
Ethan Scott
Brian Davis
Trent Davidson
Jessica Buzzelli

Term Project
ISyE 3039 – Methods for Quality Improvement – Fall 2019

Table of Contents

1. Introduction

Animal shelters struggle to place cats and dogs into permanent homes at the same rate at which animals are received. Since non-euthansia animal shelters have no say in the "quality" of the animals they receive and care for, the team explored opportunities that could improve shelters' operations while lower-demand animals continue to arrive into the system. In this project, we sought out to create a model to identify animals predicted to be adopted quickly, or those with high "adoptability," such that shelters could assign them higher adoption fees.

Using data from Petfinder.com[1], the team began by constructing a dataset of shelter animals' characteristics and the speeds at which they were adopted (refer to Appendix A for summary statistics). In the Analyze phase, we then created a variety of models in order to quantify the relationship between the characteristics and animals' time until adoption.

Though our preliminary analysis supported our hypothesis that animals' time until adoption was the result of a linear relationship between the factors, we ultimately concluded that a regression model is not an adequate representation of the data.

After training the Naïve Bayes model, we then transformed the model's outputs in order to assign adoption fees to the animals based on their predicted time until adoption (see 2: Problem Description). Unfortunately, we would recommend against the use of our model due to the classifier's high margin of error. For future work, we advise beginning with further preliminary analysis to validate our core assumptions and using a different classification method such as a neural network with non-linear activation or spatial clustering.

Table 1. DMAIC Framework

DMAIC Phase	Project Element
Define	Project proposal and background research into animal shelter operations
Measure	Collect and reshape data containing animals' individual characteristics and adoption status from Petfinder.com
Analyze	Conduct regression analysis to isolate significant coefficients for modeling adoptability; train a random forest classifier
Improve	Extend regression models to assign adoption fees within a preset range according to adoptability score
Control	Sample characteristics of recently acquired animals and determine expected incremental revenue due to new fee structure

2. Problem Description

Petfinder is a website that hosts animal shelters' listings to help match adoptable animals to potential owners. Per the site's policy, shelters can elect to charge adoption fees to help cover the operational costs and additional services such as vaccinations and sterilizations.

[1] Authentication tokens to access raw data were obtained from petfinder.com/developers/.

Petfinder has identified that animals not adopted after three months are likely struggle to become adopted at any point in the future and suggests pricing each animal between $0-$400 based on the shelter's best guess of the animal's demand. Usually, this method results in shelters applying fees of $0, $100, $200, $300, and $400 based on one of five general timespans until adoption they expect the animal to take (see 3: Data and Preliminary analysis for a more detailed explanation of "time until adoption" data).

Since adoption fees are manually assigned, the team identified an opportunity to improve operations by building a regression model to automatically determine adoption fees, assuming our time until adoption predictions are equally or more accurate than the human volunteers.

In short, the team made the following simplifying assumptions in the Define phase:

- The randomly-pulled dataset observations are a representative sample of the site
- Only the characteristics listed on Petfinder postings can impact time until adoption
- Animals priced at one of Petfinder's recommended fees are attempting to follow its fee assignment methodology (excludes the possibility of choosing the fee by chance)
- Animals' characteristics are up-to-date and do not change while the animal is in a shelter

Our hypotheses leading into the Analyze phase included:

- Time until adoption follows a non-normal distribution distribution where most animals are adopted after longer periods of time or not at all
- Animals' time until adoption is the result of a linear combination of their characteristics
- Moving to a statistically-formulated fee assignment process will lead to more precise pricing than can be done by a human and result in higher adoption revenue over time by exploiting patterns in adoption demand

3. Data and Preliminary Analysis

The preprocessed training dataset contained 15 features and 15,000 observations scraped from Petfinder.com. Refer to Appendix A for visualizations of the sample observations' overall demographics.

In Table 2, Time Until Adoption is considered a non-numerical feature since the data was only accessible as a text field indicating a general timespan (hereon referred to as their "time until adoption bucket"): "same day," "between 2-7 days," "between 1 week and 1 month," "between 2-3 months," or "after 3 months."[2] This unexpectedly complicated our planned approach since regression outputs continuous variables and ours was provided as discrete.

Furthermore, while technically a feature, Adoption Fee was not used in any models. Since our model is assigning each animal's fee as if it just arrived to a shelter, we pulled the first

[2] Petfinder stores its data with these timeframe cutoffs. The "after 3 months" bucket includes unadopted animals.

Adoption Fee associated with each animal instead of its final fee at the time of adoption. This feature will be used later to test if our model improves upon the current pricing method.

Table 2. Dataset Features by Field Type

Numerical or True/False Features	Non-Numerical Features
Age (years)	Species
Number of Pictures on Petfinder	Sex
Number of Videos on Petfinder	Maturity Size (small, medium, large, extra large)
Sterilized	Coat Color
Dewormed	Breed
Vaccinated	Fur Length
Adoption Fee	Health (healthy, minor injury, serious injury)
	Time Until Adoption*

We began testing our hypotheses by manually identifying features that contained subgroups with varying Time Until Adoption distributions. Shown in Appendix B, factors such as species, vaccination status, and sterilization status contained two or more subgroups where time until adoption's distribution appeared among the subgroups. We interpreted the uneven distributions to be potential evidence that the presence of those features implied a relationship between animals' characteristics and their resulting adoptability. The overall time until adoption distribution is shown in Figure 1.

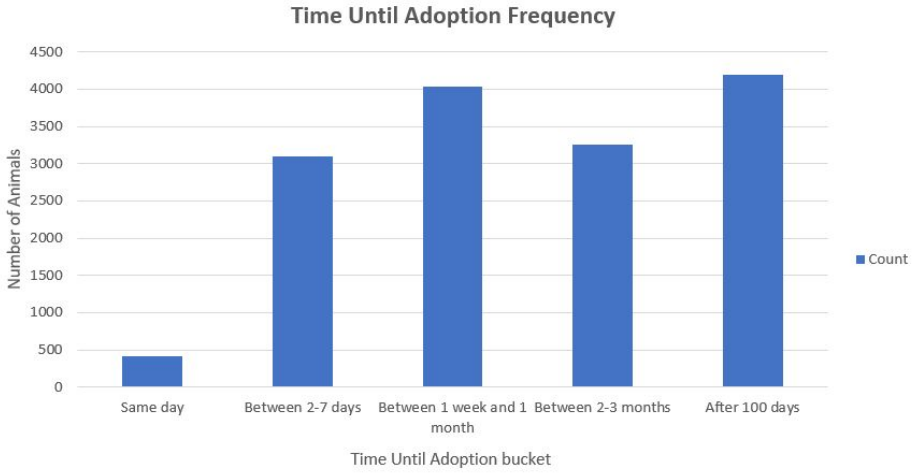

Figure 1
Time Until Adoption across all observations is non-normally, non-uniformly distributed.

After flagging potentially significant features, we transformed the dataset by one-hot encoding[3] non-numerical and true/false features and replacing observations where one or more

[3] One-hot encoding entailed taking a category containing multiple, discrete fields and replacing it with features representing each field with a 0 or 1. Example: the raw data contained many different breeds where one breed might be "Labrador / Poodle mix". We programmed a pipeline to extract breed names like "Labrador" and "Poodle" to build a set of all the unique breeds found in the data and replaced the breed feature with one feature for each breed.

features were listed as "Unknown" or blank (0.1% of total records). The resulting dataset contained 327 features and 15,000 observations.

4. Methodology and Design Solution

At the time of our project proposal, we were under the impression that Petfinder expressed time until adoption data as a continuous variable representing the number of days the animal lived in the shelter. While this was the case for a now-retired version of Petfinder's API, we found ourselves stuck with discrete labels for our model's dependent variable.

Since regression models output a continuous variable, we saw two paths forward:

1. Convert a regression model's prediction back into an integer for the "real" prediction
2. Use a classification model to predict a discrete label

Though not discussed in this course, we hypothesized a classifier would likely be more accurate than any modified regression model and thus chose to investigate both techniques.

I. Continuous to Discrete Transform Approach

Since the time until adoption data came pre-grouped into five buckets, we decided to start with an ordinary least squares (OLS) linear regression model that mapped the features to a number representing the time until adoption bucket of the observed animal. Excluding insignificant features, we arrived at the output table shown in Figure 2 and tentatively accepted our hypothesis that the data follows a linear relationship since the R^2 value was relatively large with nine significant coefficients. This hypothesis was later further upheld by a logistic regression model that returned a pseudo-R^2 value of 0.30.

OLS Regression Results

Dep. Variable:	AdoptionSpeed	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	7321.
Date:	Sun, 01 Dec 2019	Prob (F-statistic):	0.00
Time:	12:14:14	Log-Likelihood:	-23969.
No. Observations:	14999	AIC:	4.796e+04
Df Residuals:	14990	BIC:	4.802e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Type	0.0855	0.019	4.580	0.000	0.049	0.122
Age (Years)	0.0758	0.007	11.460	0.000	0.063	0.089
Gender	0.2846	0.014	20.761	0.000	0.258	0.311
MaturitySize	0.3669	0.016	22.927	0.000	0.336	0.398
FurLength	-0.1001	0.016	-6.147	0.000	-0.132	-0.068
Vaccinated	-0.0720	0.022	-3.248	0.001	-0.115	-0.029
Dewormed	0.1026	0.021	4.965	0.000	0.062	0.143
Sterilized	0.0520	0.019	2.706	0.007	0.014	0.090
Health	1.0906	0.041	26.802	0.000	1.011	1.170

Figure 2

A linear regression captured more than 80.15% of the variation between features in its time until adoption bucket prediction.

However, while the model captured a large share of the variation between features, the model's train-test accuracy was an underwhelming 26% when the model's prediction was

rounded off to the nearest accepted integer.[4] Given the five possible outputs, the model's performance was still in the realm of random chance.

Due to the model's low accuracy, we explored several refinement methods without success. First, we noticed no significant improvements by normalizing the data with a Yeo-Johnson[5] transformation or by one-hot encoding observations' time until adoption buckets for use in a multivariate linear regression. We were also unable to fit a more competitive model using coat color and breed data. Figure 6 in Appendix C is an example of a single linear regression trained with principal component analysis (PCA)[6] transforms on breed and coat color where only one PCA feature was deemed significant.

Not quite ready to give up on regression, we then decided to test a random forest model[7] with the expectation that it would confirm our prior findings were the best to be achieved or would be a superior predictor since it operates as an ensemble model. Although we expected the random forest to result in lower train-test accuracy than the original model due to its internal mechanisms to fight overfitting, it disappointed with a post-rounding accuracy of 12.56%. Refer to Table 3 in Appendix C for the model's significance scores.

II. Discrete Classifier Approach

Due to time constraints and our unfamiliarity with the subject, our first and only classification model tested was a Naïve Bayes classifier.[8] The model is typically overlooked since it has a strong tendency to overfit data and has limited ability to predict data points not represented in the training sample, but we gave it a shot since we had no shortage of discrete training data. When tested on random 2:1 train-test splits from the dataset, model accuracy converged to approximately 30% and we interpreted the outcome to support the existence of a nontrivial relationship between the features.

---

[4] After rounding to the nearest integer, we then rounded each value to the nearest integer in {0,1,2,3,4} to align with the data's discrete time until adoption labels.

[5] A Yeo-Johnson transformation is an extension of the Box-Cox transformation method that allows for zeros in the untransformed data.

[6] Principal component analysis (PCA) is a technique that reduces the number of features in a dataset by combining features with high covariances such that data loss is minimized.

[7] Whereas we were manually removing many insignificant features to repeatedly test the prior models, random forests automatically generate multiple regression models by randomly assigning subsets of features to separate "decision trees."

[8] A Naïve Bayes classifier works by evaluating each observation and its associated features and updating a prior distribution assembled through the multiplicative conditional probabilities of each associated feature being associated with the desired output.

Using the Naïve Bayes classifier to predict a test set's time until adoption buckets, we then assigned fees from $0 to $400 in $100 increments to each bucket such that the model mimicked Petfinder's pricing recommendation.

5. Recommendations

In the predictive models, we ignored the feature containing the adoption fees charged for each animal. Using K-fold cross-validation[9] with K = 10 to create train-test splits, we created 10 replications (included in Appendix D) of our model's and animal shelters' accuracy when it came to assigning the correct fee based on animals' actual time until adoption.

Afterward, we conducted a t-distribution hypothesis test[10] on the cross-validation data to confirm our model's average accuracy is not equal to or greater than the shelters'. Due to this finding, we reject our hypothesis that our model constitutes an improvement on the current pricing methodology.

6. Conclusion

In this project, the magnitude of the assumptions we had to make in order to continue with our design solution was the largest cause of failure. While we expect our models lost a great deal of potential accuracy due to time until adoption data being represented as a discrete variable, our assumptions were both too generous ("only the characteristics listed on Petfinder postings can impact time until adoption") and insufficient (ignores likely possibility that animal shelters initially price low demand animals high in order to encourage demand with reduced fees, presumes all shelters have the same accuracy).

Given that Petfinder is the largest conglomerator of such data, more descriptive data at the scale used in this study seems unlikely and we recommend future efforts begin by investigating our initial claims rather than focusing on obtaining more specific data on animals' time until adoption. A pure classification model such as neural networks or spatial clustering should lead to improved results, but we hypothesize that not much can be done for obtaining accuracy in this setting due to the nature of our guiding assumptions.

7. References

Data courtesy of Petfinder.com. Used with permission.

[9] K-fold cross validation entails randomly shuffling the order of the dataset's rows and splitting observations into K number of folds. After, the model is tested on each fold once while training the model on the other K - 1 folds. This validation technique is highly efficient in this case since it does not require forming new dataset(s). K = 10 kept the folds large enough that Naive Bayes could reach ~30% accuracy through overfitting while not doing so excessively.
[10] This case merited a t-test since both populations' mean and variance is unknown and we did not choose to assume population variances are equal.

Appendix A. Dataset Summary Charts

Our dataset was built by pulling random records such that we have no reason to reject the assumption that its demographics do not reflect the population of all animals listed on Petfinder. Since the set did not contain enough samples of each different breed or age, we hypothesized that these features would be insignificant to our models.

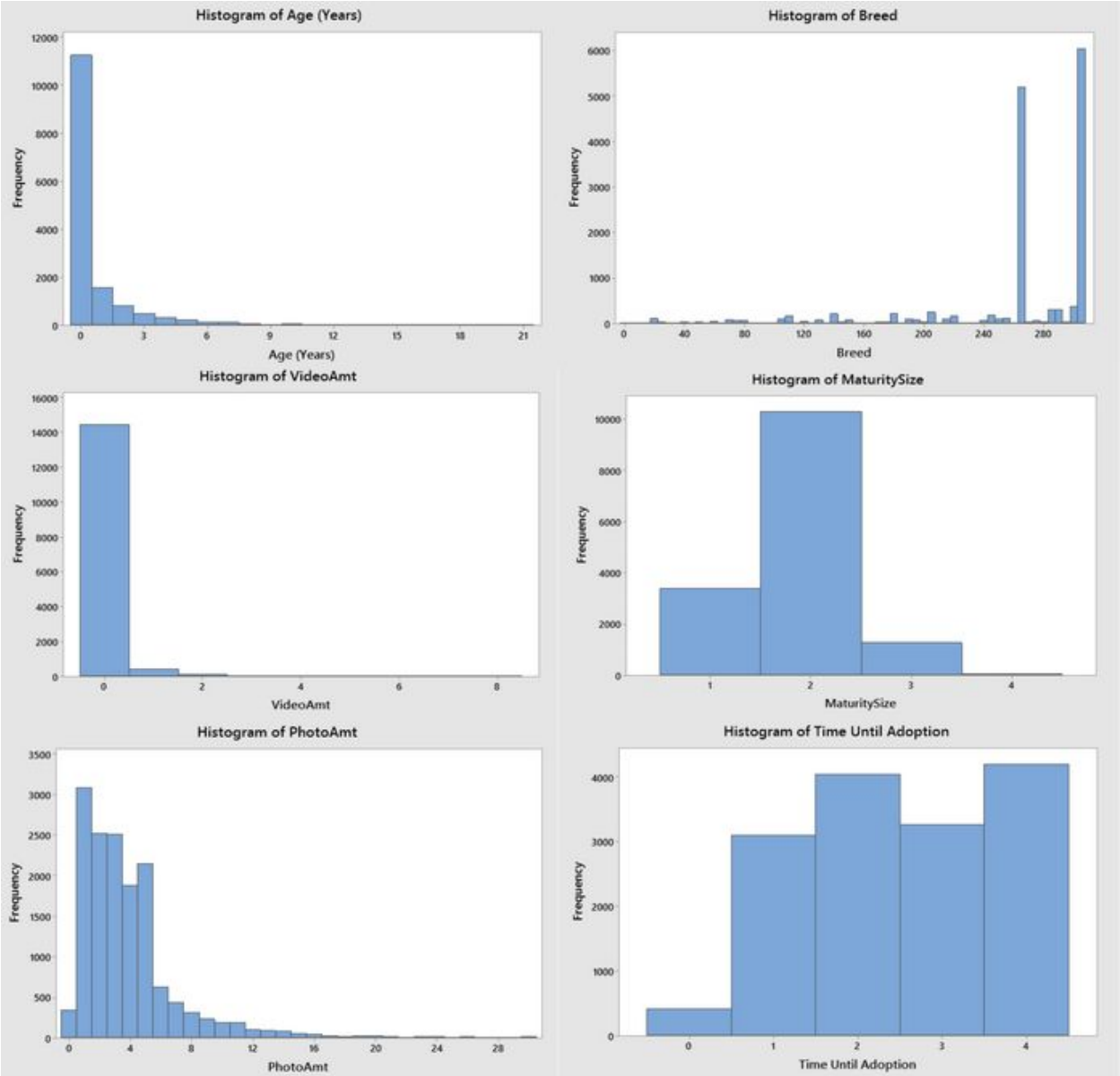

Figure 3

Dataset features were not distributed normally or uniformly. In general, our sample was mostly comprised of medium-sized, young animals that took between 1 week and 2 months to become adopted.

Appendix B. Preliminary Analysis: Proof of Highly-Correlated Features

Below is a collection of charts illustrating correlations between select features and different time until adoption buckets. The charts were created during the preliminary analysis stage and appear to support the team's belief that some features carry significant weight in defining the relationship between animals' characteristics and their time until adoption.

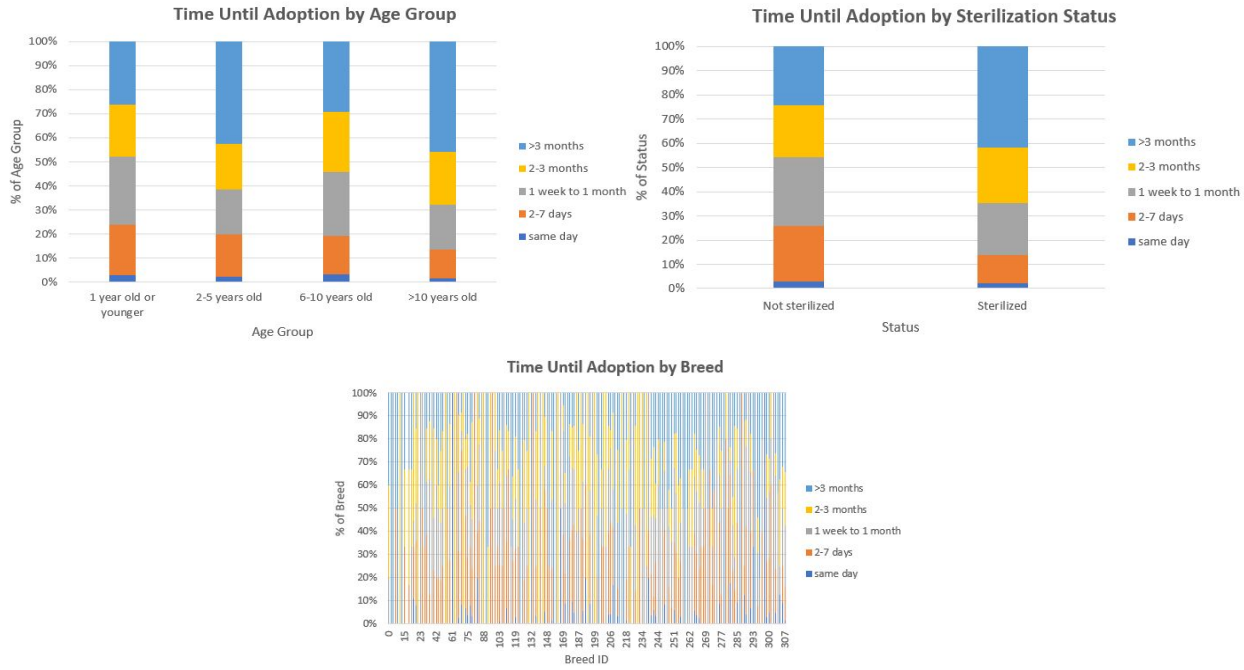

Figure 4

Time Until Adoption distributions vary across selected features' different subgroups, indicating the features are likely to cause differences in adoptability between animals.

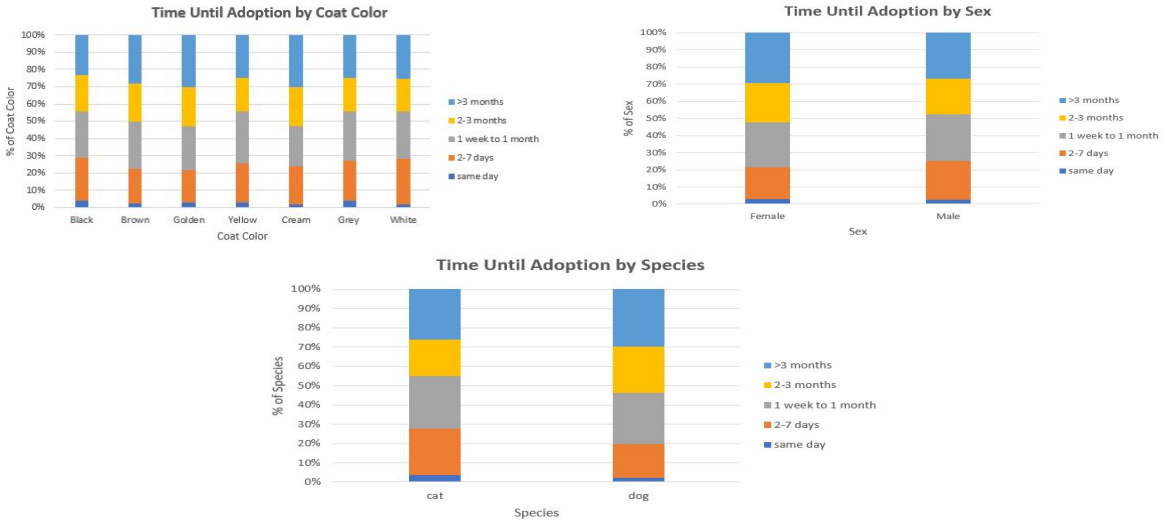

Figure 5

Compared to Figure 4, the distribution of adoption times appears more uniform across age groups and different coat colors. Slight variances between subgroups' distributions likely caused by differences in subgroups' sample sizes.

Appendix C. Regression Model Supporting Analysis

The following tables are outputs from various alternative models described in 4: Methodology and Design Solution. Figure 6 shows breed and color features are insignificant to the linear regression model. Table 3 lists the significance scores of each feature included in a random forest model where breed, coat color, sterilization status, and number of videos on Petfinder are all considered insignificant.

OLS Regression Results

Dep. Variable:	AdoptionSpeed	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.814
Method:	Least Squares	F-statistic:	2740.
Date:	Sun, 01 Dec 2019	Prob (F-statistic):	0.00
Time:	14:10:56	Log-Likelihood:	-23974.
No. Observations:	14999	AIC:	4.800e+04
Df Residuals:	14975	BIC:	4.818e+04
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Age (Years)	0.0275	0.002	11.444	0.000	0.023	0.032
Gender	0.0598	0.005	11.890	0.000	0.050	0.070
MaturitySize	0.0492	0.006	8.345	0.000	0.038	0.061
FurLength	-0.0436	0.006	-7.364	0.000	-0.055	-0.032
Vaccinated	-0.0440	0.008	-5.637	0.000	-0.059	-0.029
Dewormed	0.0496	0.008	6.553	0.000	0.035	0.064
Health	0.1621	0.014	11.709	0.000	0.135	0.189
PhotoAmt	-0.0117	0.001	-11.199	0.000	-0.014	-0.010
PCA0	-0.0035	0.005	-0.684	0.494	-0.013	0.007
PCA1	0.0034	0.006	0.542	0.588	-0.009	0.016
PCA2	0.0150	0.007	2.092	0.036	0.001	0.029
PCA3	0.0035	0.008	0.461	0.645	-0.011	0.018
PCA4	0.0046	0.010	0.468	0.640	-0.015	0.024
PCA5	0.0119	0.011	1.045	0.296	-0.010	0.034
PCA6	0.0072	0.012	0.614	0.539	-0.016	0.030
PCA7	0.0204	0.015	1.399	0.162	-0.008	0.049
PCA8	-0.0290	0.017	-1.748	0.081	-0.061	0.004
PCA9	-0.0151	0.019	-0.792	0.428	-0.052	0.022
PCA10	0.0026	0.019	0.134	0.893	-0.036	0.041
PCA11	0.0667	0.020	3.404	0.001	0.028	0.105
PCA12	-0.0015	0.020	-0.074	0.941	-0.041	0.038
PCA13	-0.0066	0.021	-0.312	0.755	-0.048	0.035
PCA14	-0.0120	0.022	-0.551	0.582	-0.055	0.031
PCA15	-0.0085	0.023	-0.366	0.714	-0.054	0.037

Figure 6

An ANOVA table from a simple linear regression model using PCA to reduce the total number of features representing breed and coat color data from 306 to 15.

Table 3. Feature Significance in Random Forest Regressor

Feature	Significance Score (%)
Age	31.449
Number of Pictures on Petfinder	12.367
Vaccinated	18.865
Dewormed	3.418
Species	12.569
Sex	5.018
Maturity Size	9.415
Fur Length	0.06899

Appendix D. Naive Bayes Cross-Validation Results

The team conducted a 10-fold cross validation to generate sample means and variances representing our model and the dataset's prediction accuracy. In 5: Recommendation, we used a t-test to refute our initial hypothesis that the model has equal or greater accuracy than the current system's fee assignment method.

Table 4. K-Fold Cross Validation Results

Replication	Model Accuracy (%)	Shelter's Accuracy (%)
0	31.40	61.81
1	29.60	63.37
2	29.27	66.14
3	30.47	56.18
4	30.07	66.11
5	29.80	66.63
6	31.27	53.28
7	32.67	70.61
8	32.20	61.15
9	31.15	66.46