Hello! My Pet

Team 12 Consulting Group

Robert Bobkoskie, Ritu Chhikara, Li Liang, Zhe Wang

Georgia Institute of Technology

Abstract: Pet adoption speed is critical to non-profits such as PetFinder.my[1], Malaysia's leading animal welfare platform. Going beyond the business costs of feeding and sheltering animals, there are important humanitarian reasons to enhance animal welfare. Large stray animal populations put humans, especially children at risk of animal bites and attacks. Animal bites can transmit diseases to humans like rabies, but they can also cause serious physical damage and pain, especially to children[2]. Identifying attributes that increase pet adoption speed will guide shelters and rescuers to improve their pet profiles, reducing both time and cost in the adoption process, lowering euthanization, controlling stray animal population while uniting rescued animals with their prospective families for unlimited happiness.

Table of Contents

I. INTRODUCTION

A. Background & Purpose

According to the American Society for the Prevention of Cruelty to Animals (ASPCA), approximately 6.3 million companion animals enter the U.S. animal shelters each year[3]. Most of the sheltered animals are adopted, bringing love and happiness to families all around. However, approximately 920,000 sheltered animals are also euthanized each year, preventing both the animals and families from potential happiness. Shelters should maximize pet adoption speed not only to optimize resources and decrease costs but also to enhance animal welfare.

The purpose of this project are due to both business and humanitarian reasons. A more efficient adoption process can lead to better management of the stray and feral animal population. According Shelby Lloyd's attitude survey on pet ownership[4], many health and community officials were concerned about irresponsible pet ownership as well as the resulting health problems associated with stray animal population. This survey evaluated attitudes of 910 pet owners and non-owners, and 697 of 910 respondents either strongly agree or agree that stray animal overpopulation was a major problem and "free roaming" animals often led to various accidents.

Pet adoption aims to unite animals with their prospective families, and together, create abundance of love and happiness. As Elizabeth Hirschman[5] wrote in her article, published in the "Journal of Consumer Research", animals serve consumers as friends. They "may be especially valuable and comforting because they provide unconditional love and loyalty" (Hirschman 620). There are numerous anecdotes of pets being more than just animals, rather, they are "extensions of self, friends, and family" (Hirschman 623). Many of these animals are living in shelters waiting for the right family, and faster adoption speed can help unite animals with their prospective families for unlimited happiness.

B. Problem Statement & Initial Hypothesis

Therefore, our problem statement is to identify attributes associated with increasing pet adoption speed at shelters. We plan to use both Logistic Regression (LR) and K-Nearest Neighbor (KNN), both classification algorithms, in our analysis to determine what attributes improve pet adoption speed. Our initial hypothesis is that age, maturity size, and photo amount will be the most important factors in predicting pet adoption speed.

II. DATA

A. Data Description

We will utilize at least two datasets for this study. Our primary dataset "train.csv" [14993-rows, 24-columns] was sourced from Kaggle and originates from PetFinder.my. PetFinder.my has been Malaysia's leading animal welfare platform since 2008, with a database of more than 150,000 animals. We also utilized two additional datasets which will be discussed later in this section.

B. Variable Description

From our primary we've identified the following variables:

1) Response Variable: AdoptionSpeed.
2) Exploratory Variables: Type, Age, Gender, MaturitySize, FurLength, Vaccinated, Dewormed, Sterilized, Health, Fee, VideoAmt, Description, and PhotoAmt.

Response Variable, AdoptionSpeed: Categorical speed of adoption. Lower is faster.

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.

Categorical Variables:

Name	Description
Type	Type of animal (1 = Dog, 2 = Cat)
Age	Age of pet when listed, in months
Breed1	Primary breed of pet
Gender	Gender of pet (1 = Male, 2 = Female, 3 = Mixed, if profile represents group of pets)
Color	Colors of pets
MaturitySize	Size at maturity (1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large, 0 = Not Specified)
FurLength	Fur length (1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified)
Vaccinated	Pet has been vaccinated (1 = Yes, 2 = No, 3 = Not Sure)
Dewormed	Pet has been dewormed (1 = Yes, 2 = No, 3 = Not Sure)
Sterilized	Pet has been spayed / neutered (1 = Yes, 2 = No, 3 = Not Sure)
Health	Health Condition (1 = Healthy, 2 = Minor Injury, 3 = Serious Injury, 0 = Not Specified)
Fee	Adoption fee (0 = Free)

C. Data Cleaning: Primary Dataset

Our primary dataset contains both Type = "cat" and Type = "dog". During team discussion, we concluded the Type of pet is very personal and might have features in the data closely correlated to Type that might impact model performance. We concluded the models may have more predictive power if the data were filtered by Type. Thus, the primary was filtered by "Type" (1 = dog, 2 = cat) to create two new datasets one of just dogs, the other all cats. For this paper, we evaluated Type = "dog" and had a new dataset with 8128 rows.

Our response variable is "AdoptionSpeed". This attribute maps to "how quickly, if at all, a pet is adopted". This variable consisted of five categories. Since we planned to use Logistic Regression for our analysis, we needed to create a binary logistic regression model, where "AdoptionSpeed" was mapped to a new categorical variable [AdoptionSpeed_fast], where [AdoptionSpeed >3] is slow [0] and AdoptionSpeed <= 2 is fast [1]. Our new response variable was appended to the primary as a new column "adoption_speed_fast" (see Figure 1.1).

Figure 1.1: Mapping [AdoptionSpeed] to a Binary Response Varibale

AdoptionSpeed	adoption_speed_fast
3	0
3	0
3	0
1	1
1	1
2	1
2	1
4	0

Final note before we leave the data cleaning of the primary dataset.

*While parsing our primary dataset, Rscript encountered an error, "**invalid 'pattern argument**" that prevented execution.*

With a little digging, we found zero values in the [Breed1] feature.

This required additional data cleaning to remove five rows with zero values for the "Breed1" attribute.

D. Data Cleaning: Second Dataset

Upon further analysis of the features in our primary dataset, we identified a feature [Breed1] that consisted of 307 unique breeds. The high dimensionality of this variable led us to conclude the potential of overfitting existed if we used this feature in our models. To address this we decided to use the breed of dog as a segue into the second dataset that would map the 307 breeds to a scale ~[1:5] that matched the other attributes in our primary.

We found a dataset that identified a dangerous dogs by breed, Declared__Dangerous_Dogs.csv[6] (dangerous dog). This dataset [63-rows, 6-columns] identifies dangerous dogs in the city of Austin TX and Travis County TX. "*These dogs are always court ordered to be restrained and should be wearing a large tag identifying them as a dangerous dog as they have attacked in the past*". Only the fifth column [dangerous_dog], which contained a character string that labeled a dog by breed was used in our approach to reduce the dimensionality of [Breed1]. Figure 1.2 shows the fifth column and the pattern in the character string that labeled a breed of dog as dangerous.

Figure 1.2: Example of Dangerous Dog Classifier in Fifth Column

6	Deirdre	Mitchell	11824 Mor	78617	â€œLady Bugâ€ spayed female, white/black Pit bull/Jack Russell mix
7	Jill	Kolansins	5336 Magc	78704	â€œTug,â€ male, brown merle and white Queensland Heeler mix
8	Maria	Davila	4420 Dove	78744	â€œTiny,â€ male, tan and white Boxer mix
9	Richard	Ashcraft	11511 Cata	78759	â€œBumpy,â€ neutered male, white and black Bull Terrier
10	Kim	Sadler	7916 Adel	78739	"Sydney" spayed female, Tricolor/Black Beagle

Figure 1.3: Example of Dangerous Dog

Breed1	breed_name	dangerous_dog_cnt	dangerous_dog
189	Rottweiler	0	0
128	Jack Russell Terrier	0	0
213	Spitz	0	0
141	Labrador Retriever	5	1
307	Mixed Breed	0	0
173	Pit Bull Terrier	2	1
119	Husky	1	0
218	Terrier	5	1

Our algorithm identified pattern matches between breed names in the primary dataset and character strings in column five in the dangerous dog dataset. However, we first needed to reverse map [Breed1], which was an integer in our primary, to a breed name which was a character string (in a lookup table) provided within the primary. We added a new column [breed_name] and were now able to iterate through the primary and the fifth column in dangerous dogs to label a dog as dangerous [1] or not [0].

To account for outliers in the mapping, we counted the number of times a dog breed matched a character string in dangerous dogs. The counter was a threshold that we could adjust. For example, if the threshold was two and we counted more than two matches, then the dog was labeled as dangerous. Figure 1.3 shows the process: new columns and thresholding.

Notice in Figure 1.3 the breeds of dog that are labeled as dangerous. Aside from the "human factor", the clerk that providing an initial labeling of the dangerous dog in report, there is evidence that a more sophisticated algorithm was required. Note in Figure 1.3, the matches on "Jack Russel", "Beagle", "Bull Terrier" and "Terrier" as these breeds were character string in column five in dangerous dogs as well as listed as a breed in our primary. We did not match on "Pit Bull" as this breed was named "Pit Bull Terrier" in our primary. With additional time, we plan to refine the algorithm to match partial character string and refine the threshold counter to better classify dangerous dogs.

E. Data Cleaning: Third Dataset

Our third dataset was used to explore the use of our models to predict the adoption speed for shelter animals for a Mock-customer in real-world scenario. The Austin Animal Center, Austin, TX provided a dataset, Austin_Animal_Center_Stray_Map.csv[7] [145190, 12] consisting of a wide range of animals in their shelter.

Figure 1.4: [Animal Type] Feature in Third Dataset

Animal Type
Cat
Dog
Other

Our models focused on dogs, so we first filtered the third dataset to only keep "Dog". Next we needed to process this data to align features with our model. It was untidy data that contained feature we could tease out and use:

[Breed1, Age, Gender, Sterilized, Color1]

Figure 1.5 shows the third dataset, cleaned, with features aligned to suit our models. Note that given time constraints to deliver this work, we did not map the breed to a dangerous dog.

Figure 1.5: Third Dataset Cleaned with Features Alignment

Breed1	Color	Color1	Sex	Gender	Sterilized	Age	Age_in_Month
German Shepherd Mix	Brown	Brown	Spayed Female	1	1	7 years	84
Pug	Black	Black	Unknown	0	0	7 years	84
Pit Bull	Brown Brindle/White	Brown	Intact Female	1	2	8 months	8
Pit Bull	Blue/White	Blue	Intact Male	1	2	3 years	36
Black Mouth Cur	Brown/Black	Brown	Intact Female	1	2	5 weeks	1
German Shepherd Mix	Tan/Black	Tan	Intact Male	1	2	10 months	10
Miniature Poodle	Gray	Gray	Intact Female	1	2	2 years	24

[Breed1] was not modified as we had a Breed feature in our trained model. Going forward we plan to reduce the dimensionality of this feature. [Color] was split on "/" and whitespace to capture the first color string. The [Color1] column is what we used for our model. [Sex] was split on whitespace to capture the [Sterilized] and [Gender] attributes for modeling. [Age] was split on whitespace to yield the age of the pet in months. Integer division by four was used to calculate Age in months when the age in the third dataset was measured in weeks.

III. METHODOLOGY

A. Overview

With the cleaned dataset, we will run both logistic regression and k-nearest neighbors to analyze pet adoption speed. Although they are both classification models, logistic regression is parametric and k-nearest neighbors is non-parametric. Parametric models allow us to fit an exact model with the data using parameters such as coefficients and intercepts. However, in order to rely on the model form for logistic regression, we would need to assume a linear relationship between variables and the log odds. Non-parametric models does not make specific assumptions on the data relationships, but requires more data to be accurate. We will use the strengths of both models to better analyze pet adoption speed. Logistic regression has a straightforward interpretation of the model for us to better understand the data and provide recommendations to shelters. We will then use k-nearest neighbors to confirm the logistic regression output and provide additional insight into pet adoption speed.

The clean dataset was randomly split into training and testing sets, 80% and 20% respectively, for all experiments. We trained the models on the training set, adjusted the hyper-parameters and experimented with logistic regression and k-nearest neighbor models and fit on the test set to get the test performance statistics.

Our approach for quantitative analysis identified ways to minimize Type-1 error. Recall we considered branding a pet as a marketing campaign. Minimizing false-positives (FP) can reduce Type-1 error, thus reducing marketing cost. With this in mind, we:

- Increased the cutoff value of p from 0.5 to reduce false positives and increase specificity.
- Measured accuracy.
- Measured precision.

In other words, we didn't want to increase specificity at the expense of accuracy and precision.

B. Logistic Regression

We applied the logistics regression model, a supervised machine learning algorithm to model and predict the speed of adoption. We used this model as it is one of the simplest & widely used models for classification problems.

a) Assumptions

Assumptions of Logistic Regression:

- The dependent variable is binary or dichotomous: the adoption speed takes two values, 0 being the slow speed & 1 being the fast speed of adoption.
- There should be no/very little multicollinearity among the predictor/exploratory variables: the collinearity heat map covered under the knn model. Moderate correlation observed between dewormed & vaccinated.
- Logistics Regression requires large sample sizes: the larger the sample size we can rely more on the analysis.
- The independent variables should be linearly related to the log odds: Odds are the ratios of something happening to something not happening.

b) Model & Selection

The model was used to identify characteristics that could maximize accuracy while minimizing Type-1 error, i.e., false positives (FP). We ran different models, permuting the exploratory variables and threshold value. For each scenario, we generated a confusion matrix and calculated the accuracy and the Type-1 error.

Figure 2.1: Logistic Regression Models – Summary Statistics

	Model 1 (Threshold=0.5, All predictors)	*Model 2 (Threshold=0.7, All predictors)*	*Model 3 (Threshold=0.3, All predictors)*	*Model 4 (Threshold=0.5, stepwise variable selection)*
Accuracy	60.92%	55.17%	54.43%	59.96%
Sensitivity	66.71%	96.63%	20.65%	66.59%
Specificity	52.62%	8.38%	92.54%	52.49%
Type-1 Error	47.38%	91.62%	7.46%	47.51%

Figure 2.2: Logistic Regression Models – Confusion Matrices

Model 1

Predicted Value \ Actual Value	N	P
N	402	287
P	362	575

Model 2

Predicted Value \ Actual Value	N	P
N	64	29
P	700	833

Model 3

Predicted Value \ Actual Value	N	P
N	707	684
P	57	178

Model 4

Predicted Value \ Actual Value	N	P
N	401	288
P	363	574

Based on the model results, we decided to choose Model1 as it maximizes accuracy and has an acceptable type 1 error rate. Model3 does minimize the Type-1 error however it significantly reduces the true positives from 575 to 178 and accuracy dips from 60.92% to 54.43%. While from a business perspective, Type-1 error is important, it cannot be so stringent that it reduces the true positives (TP) significantly.

c) Model Interpretation

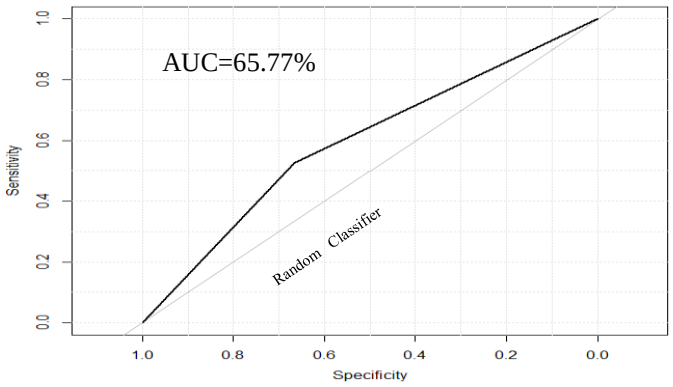
Based on the Model1 output in Figure 2.3, at 95% confidence interval Age, Breed, Maturity Size, Fur Length, Vaccinated, Sterilized are significant predictors. The negative coefficient in the output tells us that age is negatively correlated with adoption speed. Keeping all other predictors constant, the log odds of adoption speed decreases with age increasing by 1.07 units. Hence, we can deduce younger pets are adopted faster. In addition, extra-large dog breeds (MaturitySize 4) and long hair breeds (FurLength 3) have the highest coefficient within each predictor. Therefore, odds of adopting dogs with these characteristics are the highest respectively, keeping all else constant. We can also see that non-vaccinated (Vaccinated2) and non-sterilized (Sterilized2) pets are adopted faster. One possible explanation is that pet owners prefer adopting younger pets whose vaccination & sterilization are done at a later stage covered under their respective insurances. However, further research needs to be done to confirm.

Figure 2.3: Logistic Regression Model1 Output Table

Predictors	Coefficients	Std.Error	Statistic	p-value	Odds
(Intercept)	1.04274983	0.162125	6.431769	1.26E-10	2.837008
Age	-0.01079872	0.001584	-6.81618	9.35E-12	0.989259
Breed1	-0.00451807	0.00045	-10.0435	9.81E-24	0.995492
Breed2	-0.00105063	0.000218	-4.81304	1.49E-06	0.99895
Gender2	-0.3113917	0.056306	-5.53032	3.20E-08	0.732427
Gender3	-0.50905837	0.093659	-5.43522	5.47E-08	0.601061
Color12	-0.13220043	0.059454	-2.22359	0.026175895	0.876165
Color13	0.0875166	0.143039	0.611837	0.540645319	1.09146
Color14	-0.47855904	0.203704	-2.34929	0.01880938	0.619676
Color15	0.18166314	0.103049	1.762885	0.077919819	1.19921
Color16	0.1625359	0.25951	0.626319	0.531105533	1.176491
Color17	0.14243723	0.129614	1.098935	0.271796485	1.153081
MaturitySize2	-0.37955478	0.077453	-4.90043	9.56E-07	0.684166
MaturitySize3	-0.22020212	0.110701	-1.98917	0.046682722	0.802357
MaturitySize4	1.35860676	0.646871	2.100273	0.03570481	3.890769
FurLength2	0.33375049	0.055335	6.03141	1.63E-09	1.396195
FurLength3	0.71083656	0.135652	5.240131	1.60E-07	2.035694
Vaccinated2	0.36308368	0.08151	4.454488	8.41E-06	1.437756
Vaccinated3	0.35405593	0.148175	2.389447	0.016873743	1.424835
Dewormed2	-0.09054196	0.083519	-1.08408	0.278327226	0.913436
Dewormed3	-0.2795502	0.148867	-1.87785	0.060401929	0.756124
Sterilized2	0.71580374	0.073927	9.682592	3.58E-22	2.04583
Sterilized3	0.20383974	0.105629	1.929767	0.053635698	1.226102
Health2	-0.34553645	0.154039	-2.24317	0.024885757	0.707841
Health3	-0.56730434	0.510611	-1.11103	0.266555418	0.567052
dangerous_dog1	-0.02101372	0.106499	-0.19731	0.843582133	0.979206

The area under the ROC curve (AUC) provides an aggregate measure that shows how well a logistics regression model classifies all positive & negative outcomes at all possible thresholds (cut off). The larger it is better. Specific to our case, we do not have a very high threshold as it would highly decrease the TP. Hence our AUC at 65.77% is fair to the model chosen.

Figure 2.4: Logistic Regression Model1 ROC (Receiver Operating Characteristic) Curve



Residuals are the difference between what we observe and what our model predicts. As can be seen the 1Q/3Q values and Min/Max values are about to be same in absolute value and the Median is closer to 0. In addition, we also saw Min/Max values less than 3 in absolute value. This is because Deviance Residuals can be roughly approximated with a standard normal distribution.

Figure 2.5: Logistic Regression Deviance Residuals

Min	1Q	Median	3Q	Max
-2.2028	-1.0861	-0.7022	1.1413	2.2177

Figure 2.6: Logistic Regression Residual vs Fitted

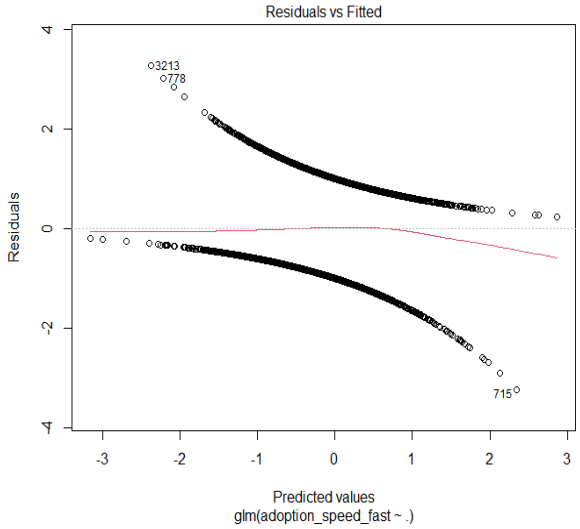
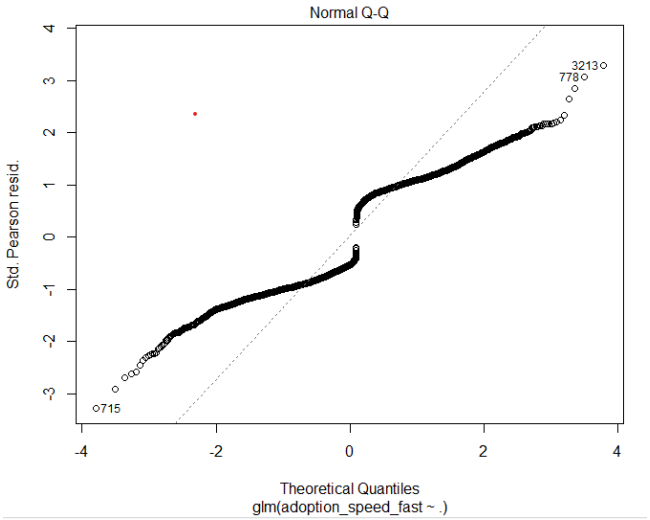


Figure 2.7: Logistic Regression Normal QQ



C. K-Nearest Neighbors

a) KNN Model Development and Interpretation

KNN classifies the data point on how its neighbor is classified. Our original target class has 5 categories from 0 to 4 with 0 referring to the highest adoption speed and 4 the lowest speed. To be consistent with multiple models development, we convert adoption speed to a 2-level binary variable, 1 for fast and 0 for slow. As knn can be used for both *binary* and *multi-class* problems, we generate two knn models on binary and multi-class targets respectively by utilizing all numerical variables to kick-start our prediction. Model based on binary response performs better with the accuracy close to 62% compared to 38% accuracy in multi-class classification.

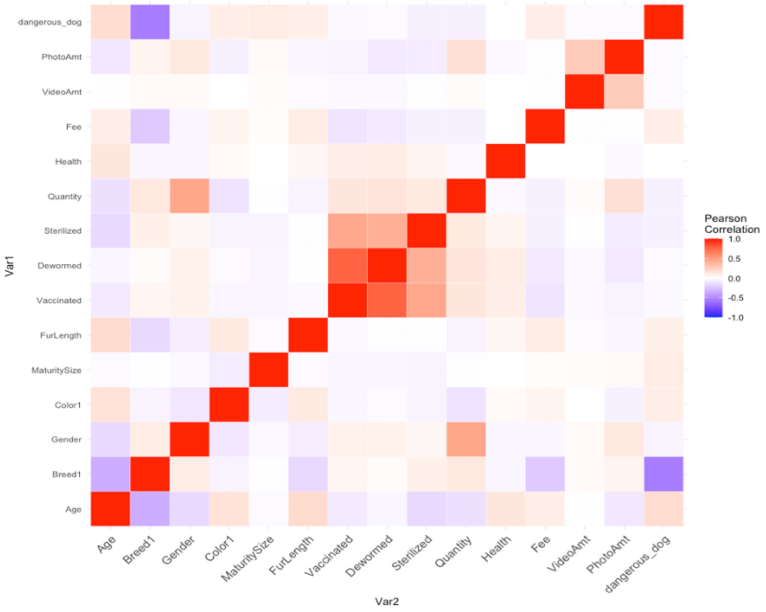
The value of k is a critical hyper-parameter in knn, but choosing k value is tricky. If k is too small, it is sensitive to noise points. Large K may include majority points from other classes, although it might work well. Tuning was accomplished by utilizing a loop to iterate k-values between [2:20]. The model produces the most accuracy with the value of k=9. Based on the result, we do feature engineering to select the most important features in order to improve model performance and reduce training time.

Figure 3.1 K-value Iteration

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Accuracy	NA	59.96	60.63	60.89	61.19	62.24	61.69	62.67	63.16	61.99
K	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Accuracy	62.67	62.73	62.55	61.99	62.3	61.87	62.73	61.93	62.12	61.75

Figure 3.2 Correlation Matrix Heat Map

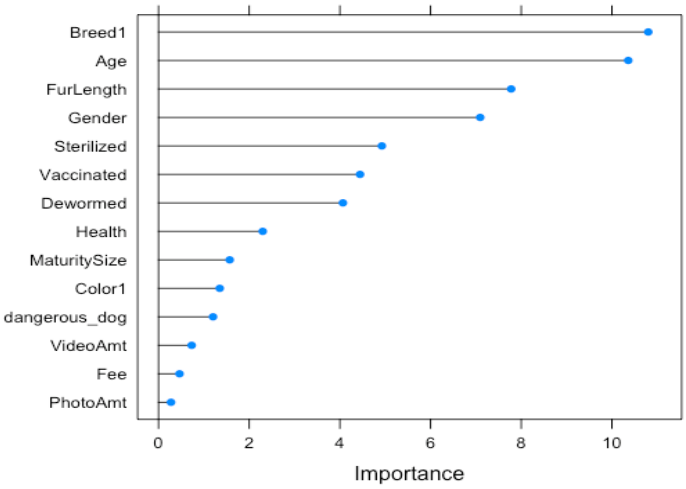
If the dataset contains features that are highly correlated with each other, the algorithm performs better if highly correlated attributes are removed. As the observation of the Correlation Matrix Heatmap (Figure 3.2), Dewormed has moderate to strong positively correlation with vaccinated. So Dewormed is removed due to its lower score in feature ranking (shown below) in order to improve model performance.



Furthermore, the importance of features can be estimated using a ROC curve analysis conducted for each attribute. In our analysis, the importance of features are estimated by building a logistic regression model. The plot indicates Breed, Age, Gender, FurLength, Sterilized, Vaccinated, Health are the top seven most important attributes. Fee and PhotoAmt are the least important.

Figure 3.3 Feature Importance Ranking

Based on the 7 most important features, knn model is developed and the performance goes up by close to 1% (Figure 3.4). Among the 7 features, Breed is a categorical variable with more than 300 categories. We convert it into dummy variable using encoding technique in order to see how well the data can be utilized to generalize the results. The introduction of dummy variables does not improve model performance but it increases the dimension and training time after adding 115 new columns. Dimension reduction techniques such as PCA usually improve the performance which we can experiment in the future



Due to the mediocre performance with long training times, we decide to remove feature Breed and replace it with dangerous dog. The performance of the new model does not see significant improvement. In model 1 and 2, k value of 2 achieves the highest accuracy, while for model 3 we obtain the most accuracy at k=20. It is observed from the table below, the specificity of the first 3 models is higher than the sensitivity, indicating that the models show higher performance when predicting the slow adoptions. This is beneficial when we try to predict pet adoptions with high length of stay. The model could help shelters identify the dogs at high risk of not being adopted soon and implement solutions to speed up the adoption, such as utilizing more resources to marketing these dogs to potential owners or relocating the pets to the shelters where they have more chances to be adopted.

Figure 3.4 Consolidated Results

		K Value	Accuracy	Sensitivity	Specificity	Type I Error
Model 1	(Feature Selection: top 7)	2	62.61%	61.65%	63.46%	36.54%
Model 2	(Dummy Variable)	2	62.12%	59.42%	64.50%	35.50%
Model 3	(Remove Breed)	20	58.92%	53.58%	63.39%	36.61%
Model 4	(Text Classification)	2	56.77%	60.45%	52.31%	47.69%

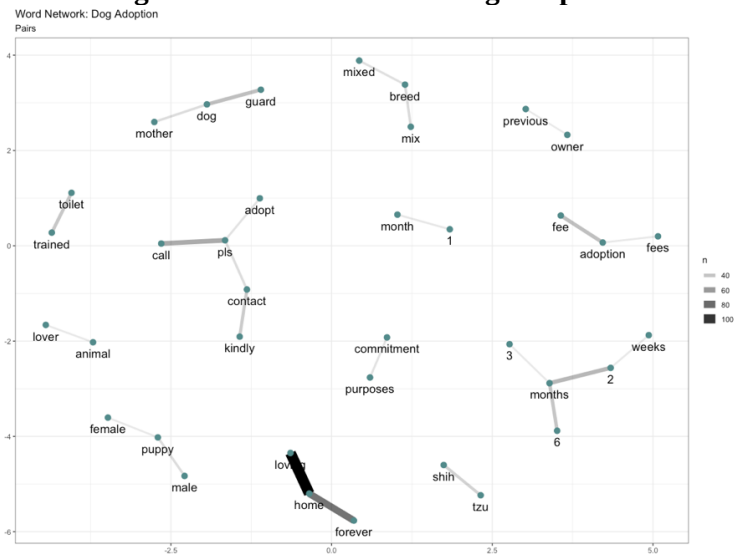
b) Text Mining

The dataset also consists of a text data: Description about the pets. Our model could be more efficient for the prediction task if dog descriptions can be incorporated into the algorithm. First we do text mining in order to discover new information or identify relationships that would otherwise get buried in the textual information.

For the dogs being adopted within 7 days, the top 50 most frequently used words in the descriptions includes home, dog, adoption, puppies, loving, playful, healthy, etc. They appear more than 200 times. It matches with what we found in logistic regression. Younger and healthy dogs are more popular.

Some words frequently appear in pairs or groups, as shown in the Figure 3.5. "Loving" and "home" appear together for around 100 times; "3", "2", "weeks" and "months" appear together for more than 40 times. This finding reinforces that feature Age is one of the most important factors influence the adoption speed. Dogs less than 6 months old will get adopted faster. The combination of "Shih" and "Tzu" appears more than 40 times, making it the most popular breed in the shelter. Shelter may consider increasing adoption fee for this specific breed and apply the fund to marketing other dogs. "Toilet" and "trained" appear together frequently, which was not discovered in our previous analysis because potty training is not a feature in the original dataset. We can conclude potty trained dogs will have more chance to get adopted quickly. We recommend shelters to do potty training and provide the information to potential owners if possible. If the shelter put some words together like "pls", "call", "contact", "kindly" in the description or "animal" and "lover", the adoption speed might go up. Further research needs to be conducted to confirm.

Figure 3.5 Word Network: Dog Adoption



As text classification can assign a set of predefined target class to open-ended text, we use knn for classifying input textual information into different categories, in our case fast and slow adoption. The model's performance is slightly lower than that of our previous knn models with the accuracy of only 57%, but sensitivity is higher than specificity. Therefore, text classification can perform a complement to our classifier model in order to predict adoption speed of fast adoptions.

IV. ANALYSIS & RECOMMENDATIONS

A. Model Result & Research Conclusion

Using best of both models (Model1 from logistic regression and Model1 from k-nearest neighbors), we can see that age, breed, fur length, sterilization, and vaccination are common important factors. Logistic regression also has maturity size as a significant factor. Both models have good accuracies, where logistic regression and k-nearest neighbors predicted 60.92% and 62.61%, respectively, of the test dataset's pet adoption speed correctly. In addition to the model accuracies, we are also interested in the Type-I error rate because it is more costly to allocate resources to dogs falsely identified as having fast adoption speed than not promoting dogs who in reality do have faster adoption speed. Despite logistic regression having slightly higher Type-I error (i.e. 47.38%) compared to k-nearest neighbor (i.e. 36.54%), we can safely use the logistic regression output to recommend action items to shelters because k-nearest neighbors confirms the logistic regression result. Therefore, to answer our research question, attributes that improve pet adoption speed are: Age, Breed, Fur Length, Vaccination, and Sterilization.

Figure 4.1: Density Plot of Predicted Probability of Response Variable

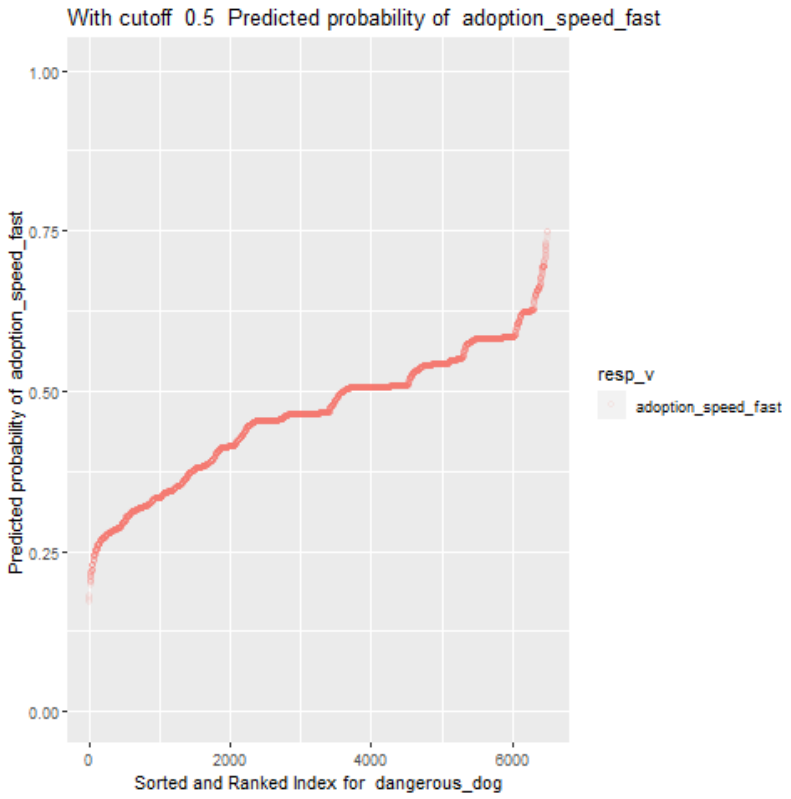


Figure 4.1 shows the predicted probability of the response variable over the sorted, ranked index for an exploratory var, in this case [dangerous_dog]. We generated these plots over several cutoff values (and k for knn) for both [dangerous_dog] and [Breed1]. However, all the plots told the same story, so we show just the one.

Figure 4.1 illustrates that our predicted probabilities are mostly centered around 0.5. The model does not predict 0 or 1 with great frequency at cutoff = 0.5.

This implies that adjusting the cutoff, lower or higher, away from 0.5 can have a very large impact on the FP, FN, TP, TN counts, thus the reported accuracy.

B. Recommendations

Our recommendation is to feature puppies and long hair dog breeds on the pet adoption home page as well as highlighting significant attributes in all pet profiles. Surprisingly, having a good health status does not improve pet adoption speed. Based on the logistic regression model, we can see that neither Health nor Dewormed are significant factors. In addition, not vaccinated and not sterilized increase pet adoption speed. Therefore, we recommend shelters to reduce funding spent on vaccination and sterilization as dogs without them have higher probability of a faster adoption speed. Moreover, shelters can also reallocate funding on deworming dogs to other areas such as rescue efforts, marketing campaigns, and local adoption drives.

V. MODEL DEPLOYMENT

The culmination of our data wrangling and modeling led us to a final test to deploy our models on a real-world dataset. As discussed in the Data section, we found a third dataset that we cleaned and could use with our logistic regression and knn models. This third dataset does not have a response variable, so we cannot generate a confusion matrix or male any claims about accuracy. However, we could train our models using features that are present in the third dataset: [Breed1, Age, Gender, Sterilized, Color1]. Recall, these feature are a subset of what was discussed in the Logistic Regression and KNN sections respectively. Below are some runs using this reduced feature set. Although we did not map the breed to a dangerous dog in the third dataset, we thought it would be interesting to compare models using either [Breed1] or [dangerous_dog], keeping all other features constant. We ran the models over varied hyperparameters and plotted results.

A. Logistic Regression: Reduced dataset [Breed1, Age, Gender, Sterilized, Color1]

[Breed1] outperforms [dangerous_dog] over most cutoff values.

- Accuracy is higher except for cutoff = 0.8.
- AUC is higher except for cutoff = 0.5.

Overall, using Breed1 outperforms, but may lead to overfitting, as discussed in section 2.D.

Figure 5.1: Logistic Regression Accuracy vs Cutoff

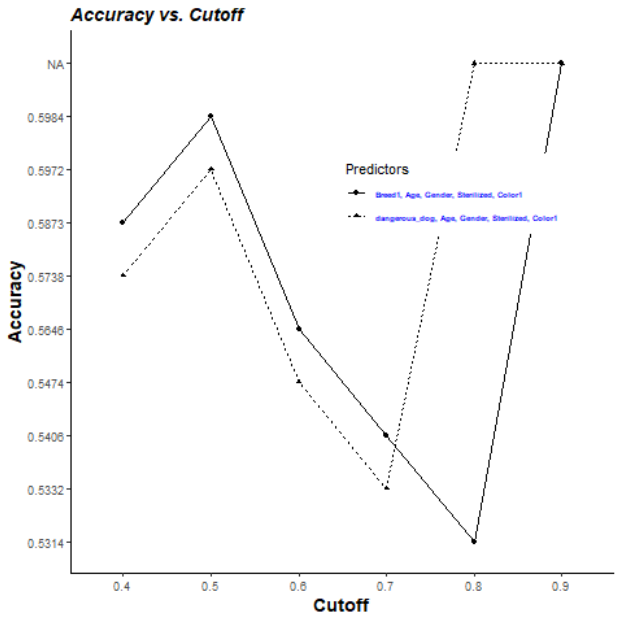
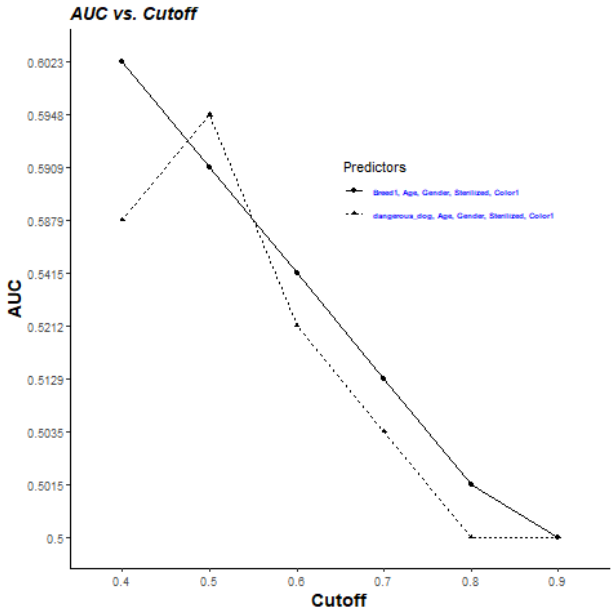


Figure 5.2: Logistic Regression AUC vs Cutoff



B. KNN: Reduced dataset [Breed1, Age, Gender, Sterilized, Color1]

[Breed1] significantly outperforms [dangerous_dog] over all k values.

- Accuracy is higher over all k values.
- False positives are lower, while providing a higher accuracy over all k values.

Overall, our KNN model tended to perform slightly better than logistic regression. As noted with Logistic Regression, using Breed1 outperforms, but may lead to overfitting, as discussed in section 2.D.

Figure 5.3: KNN Accuracy vs K Value

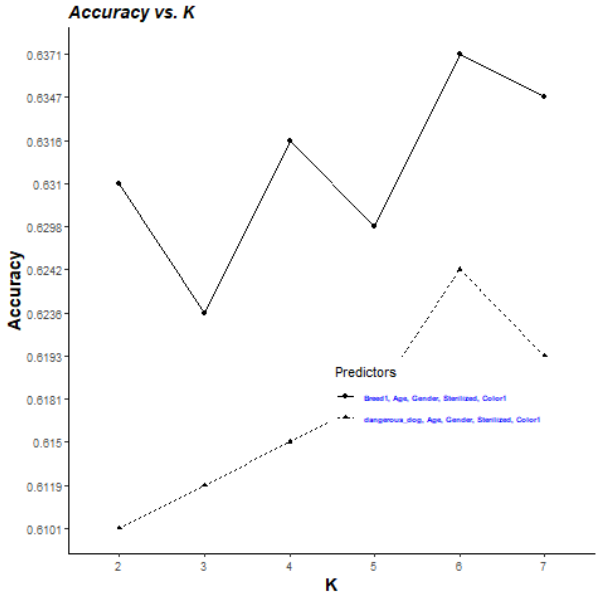
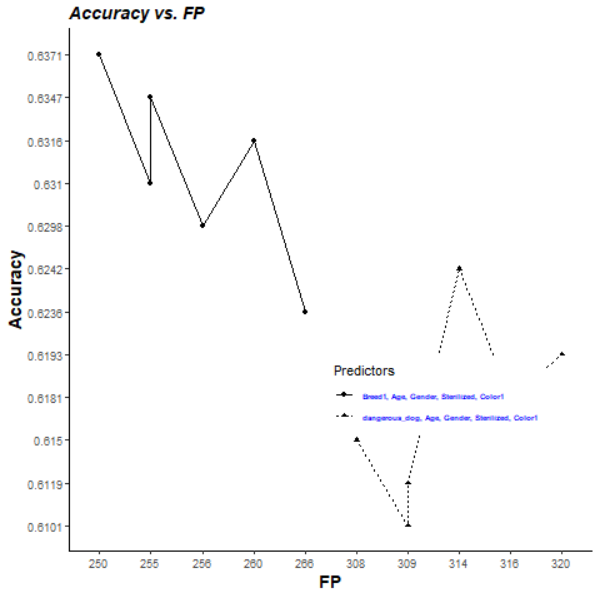


Figure 5.4: KNN Accuracy vs FP



VI. CONCLUSION

We have shown, through multiple predictive models, logistic regression and k-nearest neighbor, that identifying important attributes for increasing pet adoption speed while minimizing Type-1 error is possible. Although the feature [Breed1] in our primary dataset was found to be an important feature for predicting pet adoption speed, there large number of breeds (307) concerns us. At a minimum, this can lead to incongruity between the training and testing datasets. In other words, given a large number of potential values for a feature and a relatively small dataset in terms of rows, some feature values may not appear in either of the test/train dataset splits. In the worst-case, the model can overfit on the dataset and underperform in production.

We have attempted to address this by introducing a second dataset to correlate with [Breed1] in our primary and funnel the classification offered by [Breed1] into a reduced number of feature values. This had moderate success, but we have identified the shortfalls and can address in future work.

Our final thoughts address unifying the knowledge gained from the models into actionable business decisions for the pet shelters. We plan to optimize their advertising campaign through social media initiatives:

1. Conversion rate optimization to improve the number and quality of leads to allow the funnel to flow more efficiently and faster.
2. Develop a Conversion Rate Optimization (CRO) machine: Use surveys and web page analytics with Sale funnel analysis to rapidly iterate an advertising process to identify pet attributes that most resonate with the adoption shelter's target audience. See Appendix for a small survey we posted to our MGT6203 cohort. This was a small, sample survey to kick start the CRO machine. We will not discuss the survey in this report; however, results can be viewed in the Appendix.
3. Set up an ad campaign for the animal shelters Facebook page:
    a. Ad sets to target audience (age, gender) based on analytics, surveys and A/B testing.
    b. Objective:
        i. Consideration: traffic splits (A/B) to identify features that resonate with customer.
        ii. Conversion: sales, increase adoptions from shelter.

Using both predictive modeling and a CRO machine, we can make targeted inventory decision for shelters based on area demographics. For example, a shelter in location A might be near a senior community where older pets are preferred. We could advise Shelter-A to swap their younger pets with shelters where younger pets are in demand. Thus reducing costs and speeding adoption.

REFERENCES

[1] PetFinder.my (2019), *PetFinder.my Adoption Prediction: How cute is that doggy in the shelter?*, https://www.kaggle.com/competitions/petfinder-adoption-prediction/data.

[2] Li Ji, MD., et al, *Investigation of Posttraumatic Stress Disorder in Children After Animal-Induced Injury in China*, PEDIATRICS , Volume 126, Issue 2, August 2010.

[3] ASPCA, *Pet Statistics*, https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics.

[4] Selby, Lloyd A., et al, *A Survey of Attitudes toward Responsible Pet Ownership*, Public Health Reports (1974-), vol. 94, no. 4, 1979, pp. 380–86. JSTOR, http://www.jstor.org/stable/4596127.

[5] Hirschman, Elizabeth C., *Consumers and Their Animal Companions*, Journal of Consumer Research, vol. 20, no. 4, 1994, pp. 616–32. JSTOR, http://www.jstor.org/stable/2489763.

[6] Travis County Courts, *Declared Dangerous Dogs in the City of Austin and Travis County*, https://catalog.data.gov/dataset/declared-dangerous-dogs.

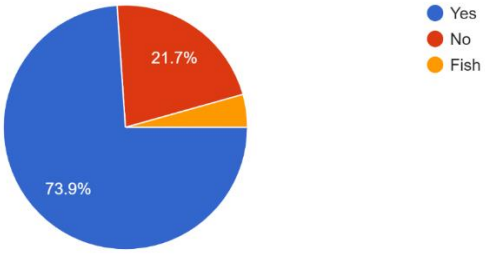
[7] data.austintexas.gov, *Austin Animal Center Stray Map*, https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Stray-Map/kz4x-q9k5.

APPENDIX

The following results are from a survey given to our MGT6203 cohort. The survey was an experiment to gather data for a CRO machine which will feed into our predictive modeling to optimize business decisions for pet shelters.

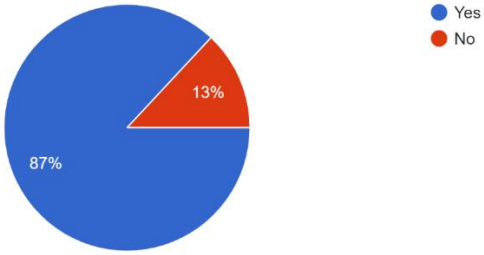
1. Have you ever been the owner of a pet cat or dog?

23 responses

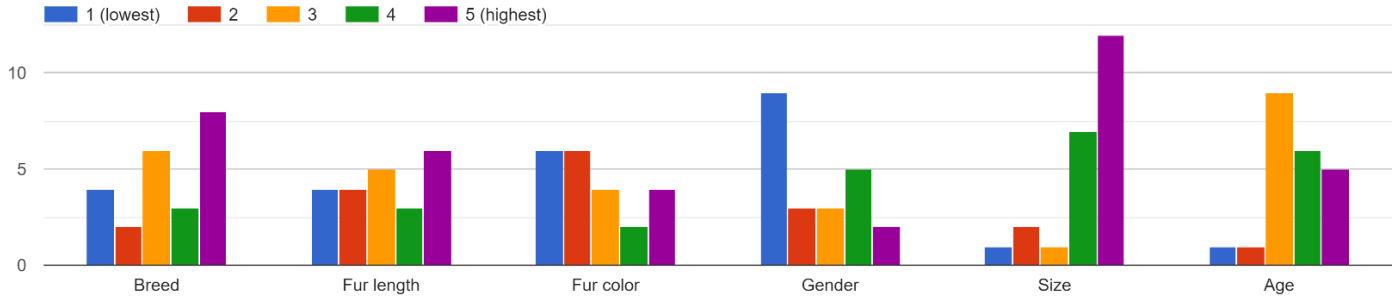


2. Have you adopted or considered adopting a pet?

23 responses



3. What are your preferred characteristics of a pet (or potential pet) ? Rank: 1 (lowest) - 5 (highest)



4. How important are these attributes in a (potential) pet? Rank: 1 (lowest) - 5 (highest)