Predicting Adoption Rates of Shelter Animals on PetFinder

Camden Olson
Princeton University
ckolson@princeton.edu

Max Piasevoli
Princeton University
mp16@princeton.edu

Abstract

Millions of animals end up in shelters every year. 1 in 4 will be euthanized. This project uses data from PetFinder in Malaysia to predict the adoption speeds of dogs and cats in addition to identifying features correlated with various adoption speeds. To this end, we train three sets of classifiers on categorical data, text data, and a combination of categorical and text data and then analyze feature importance on the fitted models. We found that a random forest classifier trained on a combination of categorical and text data best predicts adoption speed with an accuracy of 41.8%. We found that animal age, breed, and the animal's shelter were most predictive of adoption speed in addition to the use of words indicating an animal's playfulness and health.

1 Introduction

Approximately 6.5 million animals end up in U.S. shelters every year [2]. Because of crowding in shelters, almost a fourth of them will end up being euthanized [2]. With limited resources, many shelters must euthanize animals once they have stayed a certain amount of time or when the shelter becomes too full. This issue isnt unique to the U.S; countries like Malaysia also suffer from overpopulation in animal shelters [1]. To save as many animals as possible, many parties are interested in discovering what traits can improve an animals chances for adoption. Using a data set of adoption profiles and adoption speeds, we predicted animal adoption speeds. We identified factors that increase the likelihood of an animals adoption. Looking at the residuals from our predictions, we also identified factors that help an animal beat the odds and get adopted more quickly than expected. All of this information can help us determine which animals may need the most support (more promotions at adoption events, sending to a non-kill shelter, etc.) and how we can influence factors about an animal to help them get adopted more quickly.

2 Related Work

Various studies have been conducted in the past analyzing shelter data to predict adoption and identify characteristics correlated to higher adoption rates [5, 7]. One such study analyzed data from the Sacramento County Department of Animal Care and Regulation related to the adoption of 4,813 dogs and 3,301 cats from 1994-1995 and then used a multiple logistic regression to predict the likelihood of adoption [7]. The study found that age, sex, and coat color were strong indicators of adoption for both dogs and cats while a dog's breed and purebred status were also strong indicators of adoption for dogs. A second study analyzing adoption data for dogs from two no-kill animal shelters in New York observed the relation between phenotypic characteristics and the "length of stay (LOS)" at the shelter for each animal [5]. The study found that generally, LOS increased linearly with the dog's age while coat color and the sex of the dog had no correlation with LOS.

3 Methods

3.1 Data

We used a Kaggle dataset comprised of 14,998 Petfinder profiles of shelter dogs and cats in Malaysia [1]. Petfinder is an animal adoption cite where shelters post profiles for animals in need of homes. The data set include both categorical variables which have been factorized as well as text data that appear with the online adoption listings. Variables include: type of animal, age, breed(s), gender, color(s), maturity size, fur length, vaccination status, dewormed status, sterilized status, health condition, number of pets represented in profile, state in Malaysia, adoption fee, and description. We predict labels corresponding to adoption speeds (label 0 = same day listed, label 1 = 1st week listed, label 2 = 1st month, label 3 = 2nd and 3rd month, label 4 = no adoption after 100 days). The data set consists of 411 samples of label 0, 3,091 samples of label 1, 4,038 samples of label 2, 3,260 samples of label 3, and 4,198 samples of label 4. The initial data set was already divided into test and train data sets, however, the adoption speeds for the test data set were withheld. As a result, we only used the train data set of 14,998 samples and split that up into our own train and test data sets, using SciKit Learn method train_test_split, so that we could perform supervised learning methods [9]. We ensured that the distribution of class labels was evenly stratified between the train and test sets.

While using the factorized data set was possible, we chose to expand the data set using one-hot encoding. This was a more accurate way to work with the data, because it uses fewer assumptions. For example, the color of the animal was factorized as a number between 1-7 (e.g. 1 = brown, 2 = black, etc.). Methods like logistic regression would treat these values as continuous or ordinal, but not categorical. Treating these as non-categorical values does not account for the independent nature and lack of linear relationship among values in a a feature. For example, having a lot of animals with coat color 1 and another portion with coat color 7 does not mean the average animal had a color of 3.5. One-hot encoding was executed on the type of animal, breed, gender, color, maturity size, fur length, vaccination status, dewormed status, sterilized status, and state using the get_dummies function in Pandas [3]. This resulted in 374 features for the categorical data.

As for the text data, a "bag-of-words" representation was generated for each sample using the TfidfVectorizer of Scikit-Learn [9]. The TfidfVectorizer first calculates the term frequencies of all words that occur at least 10 times in the entire data set, and then re-weights the term frequencies using a TF-IDF transformation [8]. The generated vocabulary included both unigrams and bigrams resulting in a vocabulary of size 7,825. When training classifiers on solely text data, any samples without text data were removed from the training and test sets. When training classifiers on a combined representation of categorical and text data, samples without text data were generated a zero-vector bag-of-words representation.

Finally, the two data types of categorical and text data were combined resulting in a third model with 8,199 features.

3.2 Classifiers

This project utilizes the following classifiers implemented in Scikit-Learn [9]:

1. Random Forest Classifier (RF)
2. Decision Tree Classifier (DT)
3. Logistic Regression Classifier with L1-Norm (LR1)
4. Logistic Regression Classifier with L2-Norm (LR2)

In the context of multiclass classification problems, Scikit-Learn allows users to choose between estimators that are inherently multiclass classifiers and classifiers that instead use a "One-vs-Rest" approach [9]. We decided to use classifiers that are inherently multiclass classifiers in this project, so to this end for the Logistic Regression classifiers, we specified the "multi_class" parameter as "multinomial." Otherwise, the default parameters for each classifier were used. Each classifier was trained using three-fold cross-validation with recursive feature elimination cross-validation from Scikit-Learn to determine the best feature representation for each classifier [9].

3.3 Classifier Performance Metrics

To evaluate the performance of our classifiers, we consider the following performance metrics: accuracy, precision, recall, and F1-score. While accuracy simply considers the fraction of correctly predicted samples in the test set, we can define the remaining metrics as follows:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{F1} = 2(\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}})$$

Since this is a multiclass classification problem, the precision score, recall score, and F1-score were calculated for each class label and then a weighted average was taken for each statistic to account for any disparity in class frequency in the test set.

3.4 Misclassification Analysis

In addition to predicting adoption speed, we wanted to see which factors led to an animal's adoption speed being over or under-predicted. To accomplish this goal, we looked at the residuals from a Lasso regression.

We first used the LassoCV regression method in SciKit-Learn and fit the train data to the train adoption speed outcomes [9]. We then subtracted the correct adoption speed from the predicted adoption speed to calculate the residuals. We recognize that the ordinal nature of the adoption speed variable renders our data as a classification problem and not a regression problem. However, it would not make sense to take the residuals from the classification data. Lasso regression gives us a continuous variable output that provides more detail about the deviations of the predicted label from the true label. We chose LassoCV as it is a strong linear regression method that performs well with many features, as it reduces features with coefficients of zero. LassoCV also uses cross-validation to choose the best alpha parameter to fit the data.

After retrieving residuals from the train data, we used SelectKBest with F regression to determine the top 5 features important to predicting the residuals [9].

4 Results

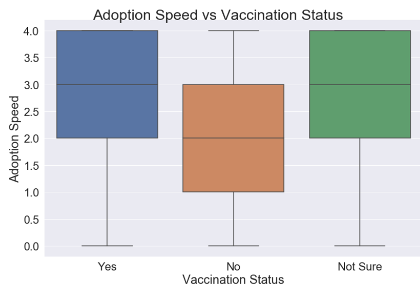
4.1 Data Exploration

We began by analyzing several features independently against adoption speed. Verifying previous research, age was positively correlated with adoption speed (Appendix: Figure 5). When looking at main color of the dog, brown animals had slower adoption rates than other colors (Appendix: Figure 6). We would have expected that darker animals had slower adoption rates, as it has been shown that darker animals are least likely to be adopted [7]. However, this finding was only somewhat supported in our analysis, since brown dogs had slower adoption rates, but black dogs did not have slower adoption rates than lighter colored dogs.

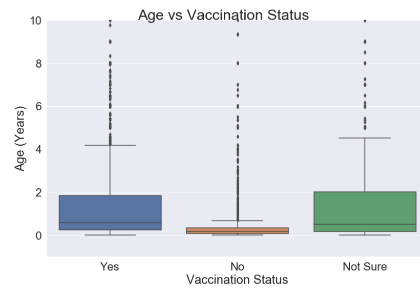
Looking at breed, there are some breeds that had faster and slowed adoption rates (Appendix: Figure 7). Spaniels and terriers had slower adoption rates. The American bulldog and Staffordshire bull terrier also had slower adoption rates. This is not entirely surprising as these breeds are considered bully breeds, a group of dog breeds that are stereotyped as aggressive because of their use in dog fighting rings. The Chinese Crested Dog was the fastest breed to get adopted, although this was a sample size of one dog. These dogs are coveted in Asian societies, and its rarity in the data set, and therefore its rarity in shelters, may also have led to its quick adoption rate. Chow chows, papillons, and dutch shepherds also had faster adoption speeds.

Interestingly, we found that animals that are not vaccinated have faster adoption speeds when compared to animals that are vaccinated (Figure 1a). However, the demographic of animals that are not vaccinated is much younger than the demographic of animals that are vaccinated (Figure 1b). Thus the faster adoption speeds for non-vaccinated animals can possibly be explained by the fact that younger animals are more likely to be adopted [5].

We also looked into unsupervised learning methods. Initially, we considered performing Principal Component Analysis (PCA) to reduce the number of features. However, a correlation matrix showed both categorical and text data had essentially no correlation (Appendix: Figure 2). Since PCA is best used with correlated features, we chose not to run any PCA analysis.



(a) A box plot of adoption speed versus vaccination status.



(b) A box plot of age in years vs vaccination status.

Figure 1: Box plots for the combined data set when grouped by vaccination status. Label 1 represents the animal has been vaccinated, label 2 represents the label has not been vaccinated, and label 3 represents the vaccination status is unknown.

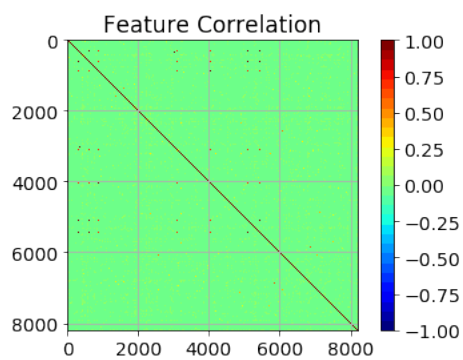


Figure 2: Correlation matrices for combined data.

4.2 Classifier Results

4.2.1 Categorical-Only Data

Diving into the classifiers, we initially trained each classifier on the categorical data alone. LR2 performed the best overall as shown in Table 1. Based on AUC scores from each class label's ROC curve, it also was the best at predicting adoption speeds 0, 1, 2 and 3 (Appendix: Table 5, Figure 8). RF was the best at predicting adoption speed 4. LR2 may have outperformed the other classifiers across all performance metrics because our features are not linearly correlated, which is an important assumption for logistic regression (Appendix: Figure 2). RF often performs better than LR [6], but perhaps only because RF does not have any underlying assumptions. Therefore, when assumptions are met, LR can perform better. RF also typically out-performs LR with more features [6]. Perhaps 374 features was small enough for LR to perform better than RF. It is not surprising that RF outperformed DT, because RF incorporates multiple decision trees, which prevents overfitting and gives a better picture of the entire data set that may not be captured in just one tree in DT [4]. DT, unlike

Classifier	RF	DT	**LR2**	LR1
Accuracy	0.368	0.344	**0.390**	0.380
Precision	0.377	0.339	**0.596**	**0.596**
Recall	0.368	0.344	**0.390**	0.380
F1-Score	0.372	0.341	**0.459**	0.454

Table 1: Performance metrics for classifiers trained on categorical data.

RF, is also unable to account for small variations in data [9]. LR2 may have out-performed LR1, because LR1 shrinks some parameters to zero. With a smaller feature set of 374 features, shrinking parameters may not be helpful and could instead disregard a fraction of the signal within the data. For example, LR1 did not output any predictions of label 0 or 1 at all, while LR2 only left out label 0. In Table 1, we also see that the logistic regression classifiers have notably high precision scores. After observing the predicted labels of the test set, we found that this high precision score is likely caused by the fact that LR1 did not output any predicted labels of class label 0 or 1. Similarly, LR2 did not output any predict label of class label 0 and very sparingly output a predicted label of classes 2 and 4.

To identify features of greatest importance for predicting adoption speed, we looked at the coefficients of each feature for each classification label fitted by LR2. Longer fur length and having a breed of poodle had the highest coefficients for predicting adoption speed 0. The fact that poodles were predictive of a fast adoption speed is not surprising. Poodles are the most popular breed in Malaysia [10]. Not being sterilized and not being vaccinated were most correlated with adoption speed 1. This is most likely because puppies are more likely to not be sterilized or vaccinated, and puppies are more likely to be adopted quickly. Photo amount was important for predicting adoption speeds 2 and 3. Age was important for predicting adoption speeds 3 and 4, which confirms previous studies showing that older dogs take longer to get adopted.

4.2.2 Text-Only Data

The performance metrics for the classifiers trained only on the text data for each sample is displayed in Table 2. We will refer to these classifiers as the ”text classifiers.” If we consider the performance metrics for the random forest classifier and decision tree classifier specifically, we see that the performance of these classifiers trained on the text data is negligibly different from the performance of the same classifiers trained on the categorical data. Instead if we consider the difference in performance between the logistic classifiers trained on the text data versus the categorical data, we can see that there is a noticeable difference in favor of the classifiers trained on categorical data. Still, it must be considered that the high precision scores of the logistic regression categorical classifiers may be a result of the fact that these classifiers only output a subset of the labels included in the data set. This restriction in prediction output also explains why the F1-score would be higher for the categorical logistic regression classifiers. Instead, the logistic regression text classifiers output all labels of the data. From this, we can infer that text data had a sufficiently strong signal-to-noise ratio for the logistic regression text classifiers to attempt to predict labels across all classes with a reasonable degree of confidence.

We also considered which words were most indicative of each class label, and to this end we considered the coefficients of the fitted Logistic Regression with L2-Norm. The top 10 most indicative words are displayed in Table 3. We see that words indicating injury like ”limping” in Label 3 or poor health like ”ill” in Label 4 are generally correlated with slower adoption speeds. Though not explicitly seen in the top 10 words of each class label, observing the most indicative words of each class label also showed that words indicating good health like ”healthy” as well as ”spayed” and ”neutered” were correlated with higher adoption speeds. Both trends intuitively make sense considering that an animal with good health will likely pose less of an initial burden to an owner upon adoption, and vice-versa.

We also plotted ROC curves for each class label and calculated the AUC scores for these ROC curves to determine the predictive abilities for each class label. ROC curves for all class labels can be viewed in Appendix Figure 9. The ROC curve for class label 4 and a table of the AUC scores are displayed in Figure 3. In Table 3b, we can see that all text classifiers predicted class label 4 very

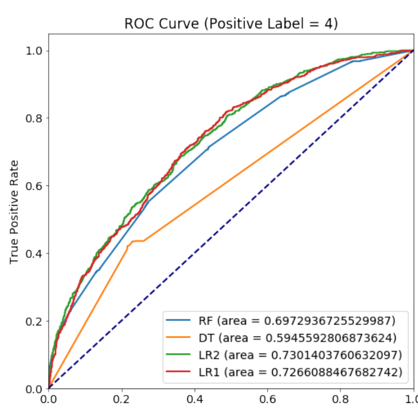
Classifier	RF	DT	**LR2**	LR1
Accuracy	0.365	0.340	**0.371**	0.363
Precision	0.38	0.343	**0.384**	0.376
Recall	0.365	0.340	**0.371**	0.363
F1-Score	0.371	0.341	**0.376**	0.368

Table 2: Performance metrics for classifiers trained on text data.

Label 0	Label 1	Label 2	Label 3	Label 4
adopted	looking	medication	best	msg
posting	already	fee rm	spaying	call give
mickey	adopt baby	picked	beauty	furry friends
active please	oversea	responsible	pet lover	ill
without mother	cleaned	must view	weight	us anyone
owners	priority given	dirty	move	adopting cute
profile	feel	whatsapp call	limping	playful also
borned	poodle	nibble	female puppy	open adoption
sleep	poor little	two cats	call home	contact furry
really love	certified	present	especially	white cat

Table 3: The 10 top most indicative words of each type of bias as determined by the Logistic Regression with L2-Norm trained only on text data.

well in relation to their respective predictive abilities of other class labels. Additionally, we see that the logistic regression classifiers predicted class label 0 very well also in relation to their predictive abilities of other class labels while the random forest classifier and decision tree classifier were not capable of predicted class label 0 as well. We can also see that all classifiers comparatively struggled to identify class labels 1, 2, and 3. These trends suggest that while there can exist word usage that helps in identifying the extreme labels of 0 and 4, there is a less distinct set of vocabulary used to describe animals of class labels 0, 1, and 2.



(a) ROC curve for text classifiers treating class label 4 as the positive label.

Classifier	RF	DT	LR2	LR1
0 Label AUC	0.639	0.544	**0.734**	**0.717**
1 Label AUC	0.613	0.554	0.631	0.626
2 Label AUC	0.571	0.558	0.578	0.569
3 Label AUC	0.603	0.548	0.610	0.600
4 Label AUC	0.697	0.595	**0.730**	**0.727**

(b) AUC scores for each text classifier across all class labels.

Figure 3: ROC curve and AUC scores for the text classifiers.

4.2.3 Categorical and Text Data Combined

The performance metrics for the classifiers trained on the combination of categorical and text data is displayed in Table 4. We will refer to these classifiers as the ”combined classifiers.” In Table 4, we see that the combined data increases performance of the random forest and decision tree classifiers across all metrics in comparison to the respective categorical classifiers and text classifiers. Conversely, the combined data decreases performance of the logistic regression classifiers across all

Classifier	**RF**	DT	LR2	LR1
Accuracy	**0.418**	0.360	0.371	0.369
Precision	**0.444**	0.359	**0.580**	**0.610**
Recall	**0.418**	0.360	0.371	0.369
F1-Score	**0.428**	0.359	0.443	0.451

Table 4: Performance metrics for classifiers trained on the combination of categorical and text data.

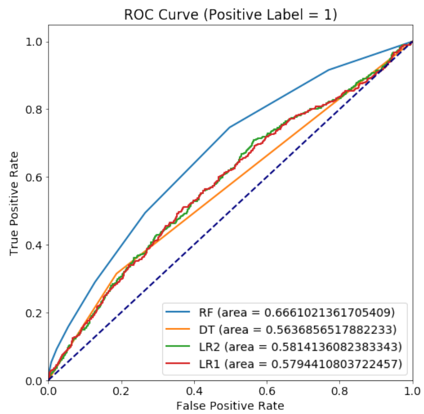
metrics in comparison to the respective categorical classifiers and text classifiers. These trends could possibly be explained by the fact that the greater feature number in the combined data caused the logistic regression classifiers to overfit to noise in the training set despite our efforts to counteract this with feature selection. Random forest and decision tree classifiers innately counteract overfitting explaining why these performance metrics did not decrease in the case of these classifiers. Based on its performance improvement when trained on the combined data, we can infer that the decision tree classifier took advantage of the combined signal of categorical and text data. The increased performance of decision tree classifiers explains the performance improvement of the random forest classifier since the random forest classifier is an ensemble classifier consisting of multiple decision trees. In Table 4, we also notice that the precision of the logistic regression classifiers is noticeably higher than the random forest and decision tree classifiers. Upon further inspection we found that the logistic regression classifiers were again only outputting predicted labels of either 2, 3, or 4 explaining why the precision scores are higher for these classifiers. Instead, the random forest and decision tree classifiers attempt to predict all labels in the data set.

We analyzed the importance of specific features in the random forest combined classifier to determine which features were most predictive of adoption speed. We found that the most important feature was the RescuerID, which is a unique identification code assigned to each shelter listed on PetFinder. We speculate that this feature can be the most important feature since many characteristics can vary between different shelters. For example, certain shelters can be located in wealthier neighborhoods which may have a greater capacity for individuals to own pets. Additionally, certain shelters may have a certain reputation in their neighborhood or greater presence on PetFinder that causes their dogs to be adopted faster. The second most important feature is age, which intuitively makes sense as past studies found that younger ages of dogs are correlated with faster adoption speeds [5]. Additionally we found that the occurrence of words like ”friendly,” ”playful,” and ”healthy” were among the most important features. Again, this trend makes sense since these are all desirable traits that prospective pet owners look for when considering whether to adopt an animal.

Again, we plotted ROC curves for each class label and calculated the AUC scores for these ROC curves to determine the predictive abilities for each class label. ROC curves for all class labels can be viewed in Appendix Figure 10. The ROC curve for class label 1 and a table of the AUC scores are displayed in Figure 4. In terms of trends in the AUC scores, we see that in Table 4b the AUC score for class label 4 is the highest among the AUC scores for each classifier. This means that each classifier has the best predictive ability for class label 4. While the decision tree and logistic regression classifiers had comparably less predictive ability for the other class labels, the random forest classifier predicted class labels 0, 1, and 4 fairly well in relation to their set of AUC scores. The random forest classifier comparatively struggled to identify class labels 2 and 3 suggesting that the combined data set does not have an adequate signal-to-noise ratio needed to predict these labels with sufficiently strong confidence.

4.3 Misclassification Analysis

We wanted to look at which features led to greater residuals and whether they were correlated with either under or over-prediction. We looked at the five most important features responsible for the residuals in the categorical-only data. The most important feature was if a dog was mixed breed. A mixed breed dog could come in any appearance, which means they probably do not have similar characteristics as a specific breed would. Therefore, it makes sense that being mixed breed would make it harder to predict that animals adoption speed. Being of medium size also made it harder to predict adoption speed. Analyzing some of the most over predicted and most under predicted animals, being medium sized led to under predictions. Sterilization status and not being vaccinated



(a) ROC curve for combined classifiers treating class label 1 as the positive label.

Classifier	RF	DT	LR2	LR1
0 Label AUC	**0.681**	0.578	0.518	0.508
1 Label AUC	**0.666**	0.564	0.581	0.579
2 Label AUC	0.606	0.549	0.596	0.587
3 Label AUC	0.617	0.554	0.595	0.591
4 Label AUC	**0.763**	0.634	0.680	0.675

(b) AUC scores for each combined classifier across all class labels.

Figure 4: ROC curve and AUC scores for the combined classifiers.

were also important for predicting residuals, but values were similar for over- and under-predicted, so there was not a clear trend of how sterilization and vaccination status led to larger margins of mis-prediction. Lack of vaccinations may have led to mis-predictions because of opposite reasons for vaccination status. As noted in the data exploration section, puppies were often not vaccinated because they are too young, but puppies get adopted more quickly. On the other hand, ill dogs or dogs at poorer shelters without resources to vaccinate could have longer adoption times.

5 Discussion and Conclusion

In this report, we first tested three sets of classifiers respectively trained on three different types of data inputs to predict adoption speeds of adoptable pets in Malaysia. We predicted adoption speeds using categorical-only, text-only, and categorical and text data combined. For categorical-only and text-only data, the logistic regression with L2-Norm performed the best across all metrics. Despite the relatively high performance statistics, we found that this classifier only predicted class labels 2, 3 and 4 explaining its noticeably high precision score. We found that the random forest classifier trained on the data set of combined data performed the best among all classifiers of any data type with an accuracy of 41.8%. We speculate that the relatively high performance statistics of the random forest classifier can be explained by its ability to prevent overfitting in a higher feature space while taking advantage of the combined signal of the categorical and text data.

This project also identified specific features that were predictive of adoption speed in each data type. Our results reaffirmed previous studies which found that adoption speed increases with the animal's age while we found that being a poodle in the case of dogs in Malaysia was correlated with higher adoption speeds. For text data, words indicating illness and injury were correlated with slower adoption speeds, and positive words like "trained," "playful" and "healthy" led to faster adoption speeds. Analyzing the feature importance of the random forest classifier in combined data set reaffirmed the correlation between age and adoption speed, but more importantly, it found that the animal's shelter had the most importance in determining adoption speed. Therefore our project has shown that there are features beyond the animal's phenotypic characteristics that can influence its adoption speed.

There are many extensions that can be applied to expand the scope of this project and improve accuracy. We recommend further feature engineering and inclusion of new features such as purebred status and level of training. Inclusion and analysis of image data may also be useful for predicting adoption speed. This data set looks at adoptable pets in Malaysia, but it would be interesting to see how trends found in this dataset translate to other countries. Additionally, we could divide this data set into dogs and cats and predict adoption speeds for each type of animal separately.

References

[1] "Petfinder.my adoption prediction," *Kaggle*, 2019, https://www.kaggle.com/c/petfinder-adoption-prediction/data.

[2] "Shelter intake and surrender," *ASPCA*, 2019, https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics.

[3] "Pandas.get_dummies," *Pandas 0.24.2 Documentation*, Accessed: 5/14/19, https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html.

[4] "What is the difference between random forest and decision trees?" *Quora*, Accessed: 5/14/19, https://www.quora.com/What-is-the-difference-between-random-forest-and-decision-trees.

[5] W. P. Brown, J. P. Davidson, and M. E. Zuefle, "Effects of phenotypic characteristics on the length of stay of dogs at two no kill animal shelters," *Journal of Applied Animal Welfare Science*, vol. 16, no. 1, pp. 2–18, 2013, pMID: 23282290. [Online]. Available: https://doi.org/10.1080/10888705.2013.740967

[6] P. P. . B. A. Couronne, R., "Random forest versus logistic regression: a large-sclae benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 270, 2018.

[7] M. Lepper, P. H. Kass, and L. A. Hart, "Prediction of adoption versus euthanasia among dogs and cats in a california animal shelter," *Journal of Applied Animal Welfare Science*, vol. 5, no. 1, pp. 29–42, 2002, pMID: 12738587. [Online]. Available: https://doi.org/10.1207/S15327604JAWS0501_3

[8] Manning, C.D., Raghavan, P., and Schütze, H., *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and e. a. O Grisel, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] C. Perro, "Popular dog breeds in malaysia," *Perro Mart*, Accessed: 5/14/19, https://perromart.com.my/blogs/perro-learning-center/popular-dog-breeds-in-malaysia.

Appendix

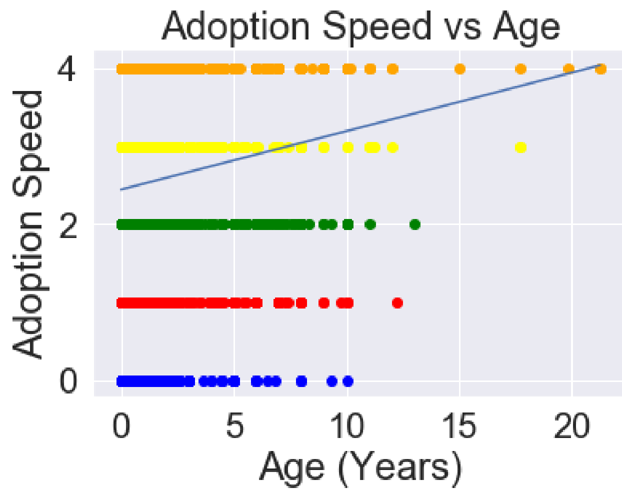


Figure 5: Plot of adoption speed versus age in years.

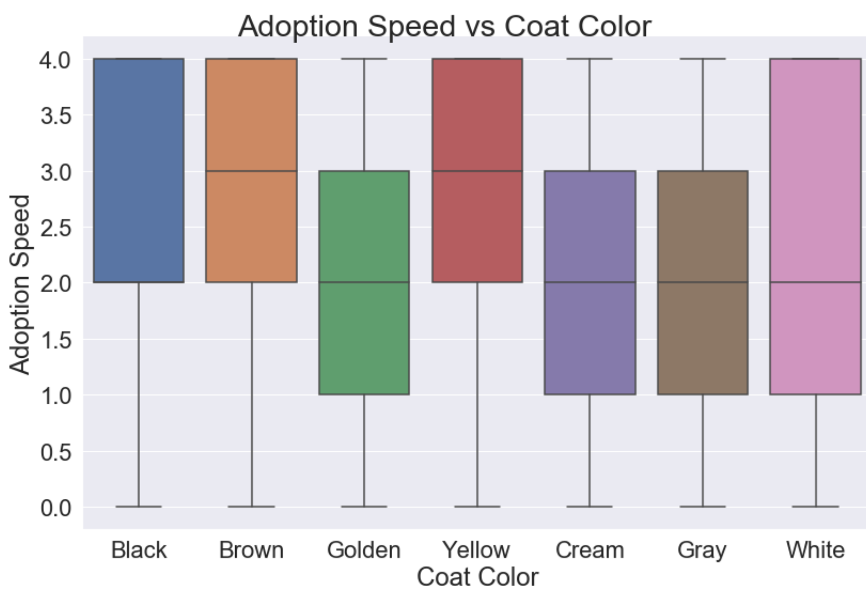


Figure 6: Box plots of adoption speed versus coat color.

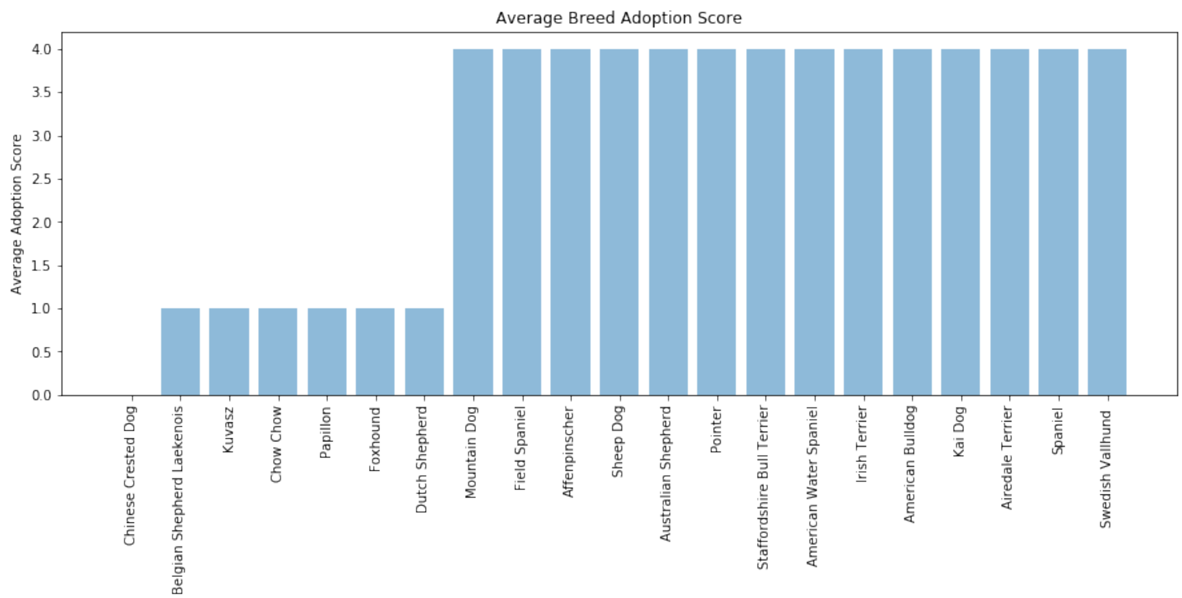


Figure 7: Average adoption speeds for dogs with fastest (0 and 1) and slowest (4) adoption speeds.

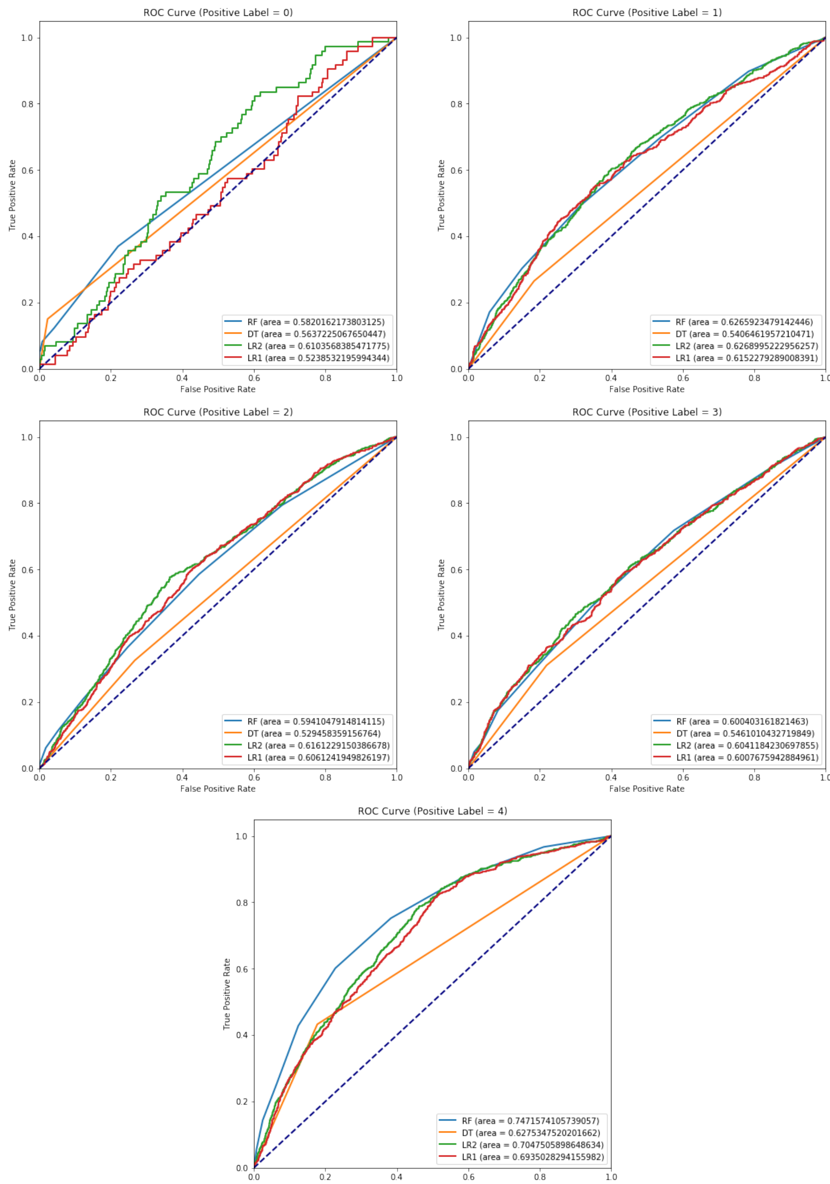


Figure 8: ROC curves for categorical data

Classifier	RF	DT	LR2	LR1
0 Label AUC	0.582	0.563	0.610	0.524
1 Label AUC	0.627	0.541	0.627	0.615
2 Label AUC	0.594	0.529	0.616	0.606
3 Label AUC	0.600	0.546	0.604	0.601
4 Label AUC	0.747	0.628	0.705	0.694

Table 5: AUC scores for categorical data.

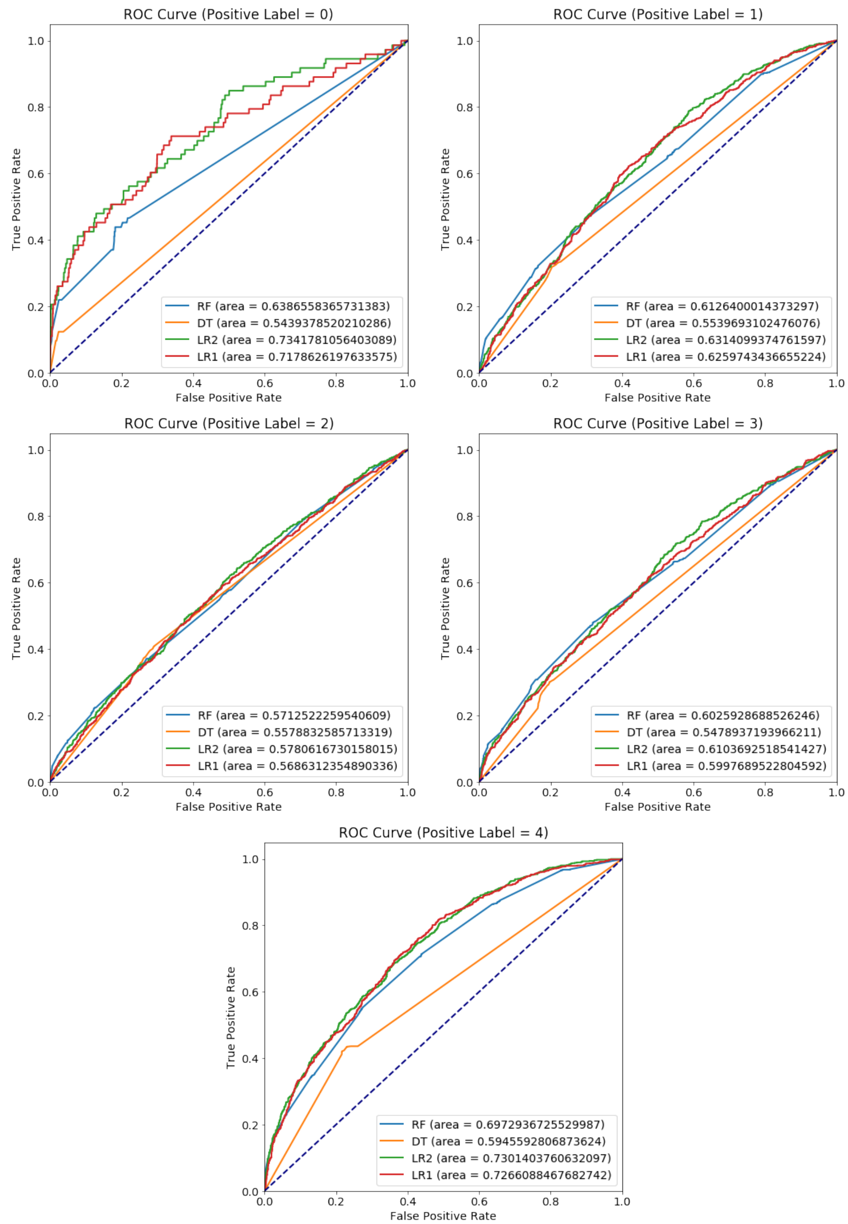


Figure 9: ROC curves for text data

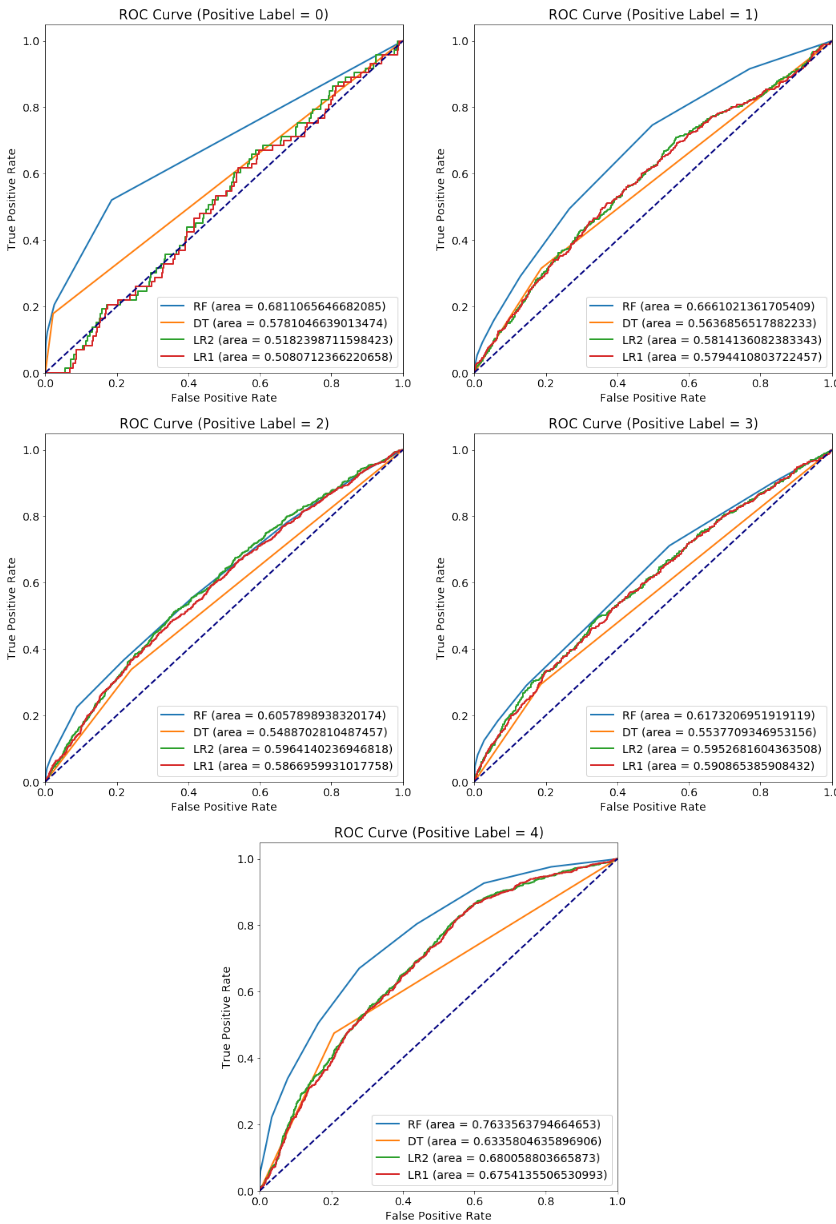


Figure 10: ROC curves for combined data.