STA 160 Final Project: Speed of Pet Adoption

Chengchen Luo, Liya Li, Ryan Kim, Sihui Li

***Abstract*-With an interest in the factors that affect pet adoption, we obtained data from Petfinder.my, a platform dedicated for pet adoptions, and applied statistical methods and tools to analyze the data. The objective is to create a model for predicting the speed at which a pet is adopted. There are two focuses in our model, one part being using the data from the text of the csv files and the other using the adopted pet image data, then comparing the results. We further looked into metadata, but it seems too difficult a task and so we briefly touch upon it here.**

I. Dataset Introduction

The original data is obtained from the link below and is named petfinder-adoption-prediction.zip: https://www.kaggle.com/c/petfinder-adoption-prediction/data and the size is about 2 GB. Our dataset contains six different sources of data including csv data, images, metadata images, and sentiment data.

II. Data Preprocessing

The csv file contains a lot of unique features about each adoption. Thus, we took a look at them in comparison to the adoption speed. Some of the features we explored and analyzed by showing graphs are animal type, gender, health of animal, age of the pet, size of the pet, color of the pet, breed, adoption fee, location by states in Malaysia, and rescuer location ID. For data cleaning, we decided not to keep the variables missing more than 50% of the observations. However, for the rest of the dataset, we will replace the numeric variables by the median and create a category called "missing" for categorical variables.

For the image data, the images zip files contained all the animal images in the format of jpg. There are multiple images that represent the same animal, and to distinguish them, image name in series are used. In train_images.zip, there are 58,311 jpg files. In test_images.zip, there are 14,465 jpg files.

III. Materials and Methods

The overall analysis in both paths is aiming to obtain prediction models: one from the pet adoption text processing and the other from the raw image data extraction from the jpg files. Then, the two models will be compared through our analysis and the prediction accuracy scores. We used Python and various packages within the language to compute all of our data analysis, in which our code can be found in 'Supp.zip.'

In the descriptions found in the csv data, we used text processing to create a Lasso model for prediction. For the image data, we converted the image data into pixel arrays and analyzed the pixels by extracting select features from the data and proceeded to feature analysis and machine learning model selection and prediction. The method to be used is built upon the idea of using neural networks for extracting features.

Furthermore, we took a look at some of the other columns found within the csv data and explored the data to help us build on drawing our analysis and understanding why we perhaps got the results that we did.

IV. Data Exploration

A. *CSV Data: Feature Analysis*

From the original analysis of the data exploration, it seems that there are a few identifiable features that have an impact on adoption speed. We have plotted 19 visualizations (but omitted some here) for comparison and conclude each feature and its relations. Our conclusion of the analysis of those features that seem significant are:

- Type: cats are adopted faster than dogs (Fig. 1)
- Mixed gender pets are adopted slower
- More fur is adopted faster
- Younger pets are adopted faster
- Small pets in size are adopted faster
- Mixed breeds are adopted faster
- Free cats seem to be adopted faster
- Certain regions (Selangor) have faster adoption rates
- Higher sentiment score means they are adopted faster
- Pet name increases adoption speed
- Longer description tended to lead to faster adoption (Fig. 2)

Features that we found to have little to no impact on the adoption speed:

- Whether the pet was vaccinated or not did not seem to matter
- Dewormed also had little to no impact
- Sterilized or not did not have much impact
- Color of the pet was insignificant

We then compiled all the features given in the csv files and metadata to summarize our visualizations in a correlation matrix seen in Fig. 3.

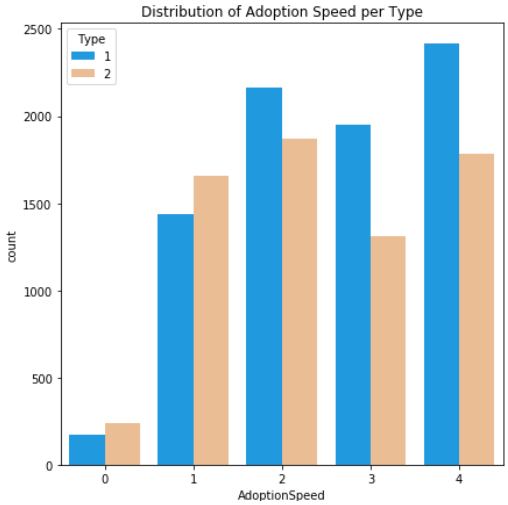

Figure 1. Adoption speed by type of pet

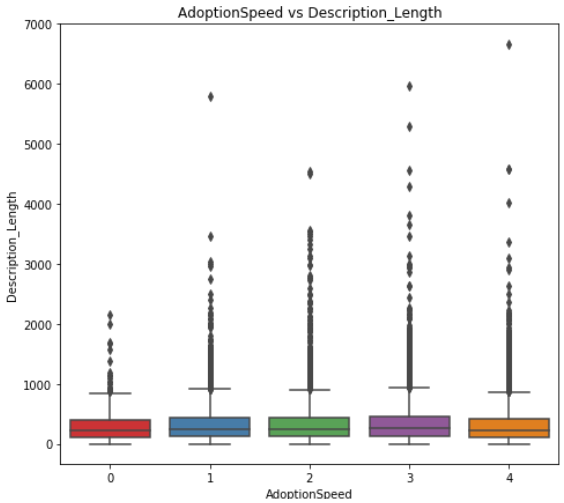

Figure 2. Adoption speed by description length

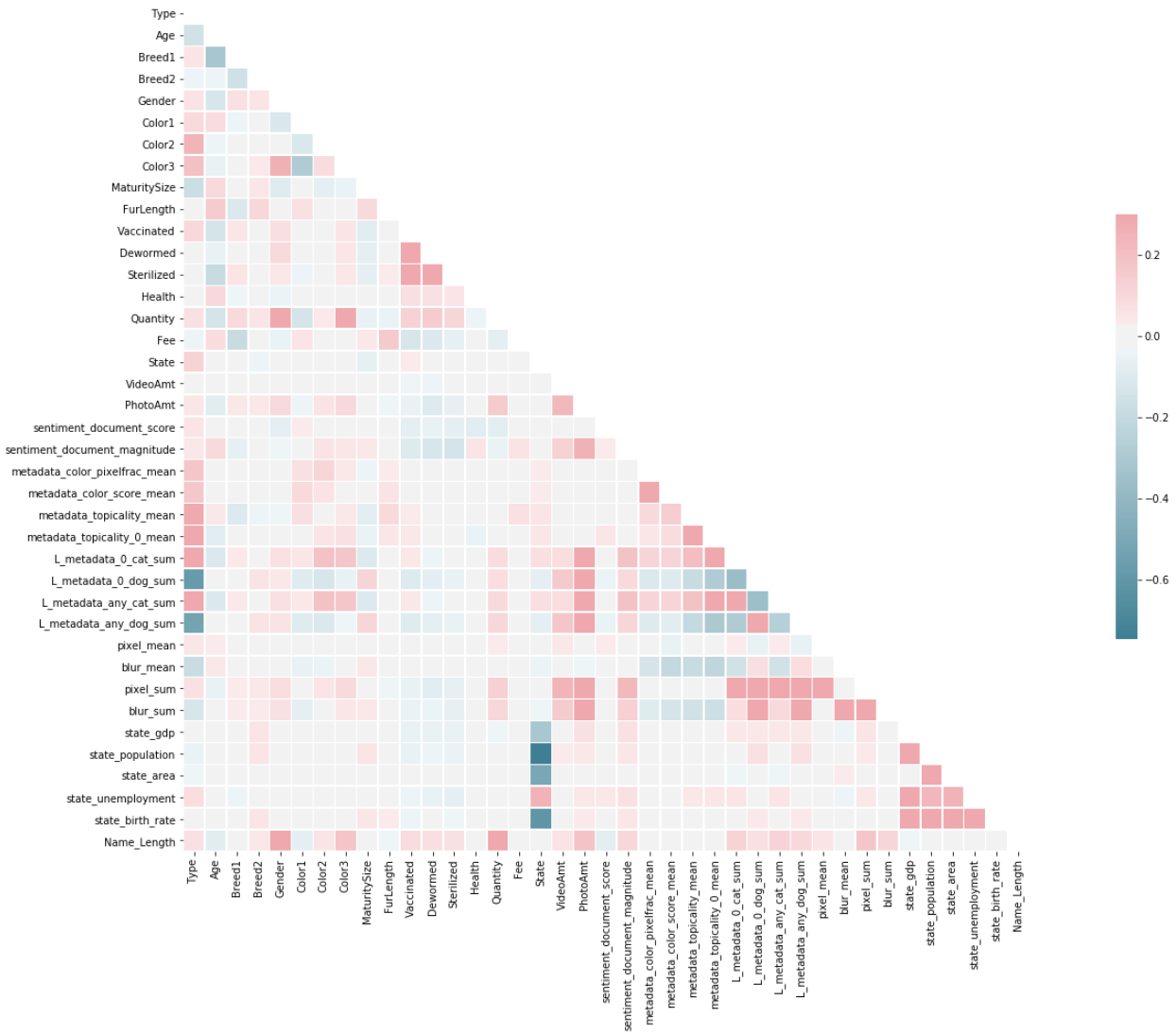

Figure 3. Correlation matrix of dataset features

B. Metadata Analysis

The metadata analysis of the description can be summarized in Fig. 4 shown below. In each plot, we use different colors to separate the group of adoption speed. As we can see, most of the density plot of the features in the descriptions are similar. From the data, the top five states with the most ads are, in descending order, Selangor (56.67%), Kuala Lumpur (27.05%), Pulau Pinang (5.68%), Johor (3.39%), and Perak (2.77%). The top three states accounted for nearly 90% of the ads. Therefore, the number of ads per state is uneven. In order to compare the adoption speed among the states, we use the metric percentage instead of frequency.

Fig. 5 shows the percentage of the observation in each adoption group per state. As we can see, Sarawak has the highest percentage in the adoption speed 4. The dataset has 18 ads posts for the state Sarawak. It means that the pets in this state usually take more than 100 days to be adopted. In addition, it appears that there is no adoption speed group 0 in the state Labuan. But we only have 7 ad posts from the state Labuan. Therefore, it is possible that we do not have

data for this group due to randomness. Selangor has the most ads with the adoption speed in group 0 to 3, which means that Selangor has the largest percentage of pets getting adopted in the first 100 days. This is also the state with the number of ads posted, with 10,547 ads posts.

Lastly, the metadata confirms that cats are more likely to be adopted early than dogs, which was shown in Fig. 1. Further, it shows that the age distributions are all skewed right.

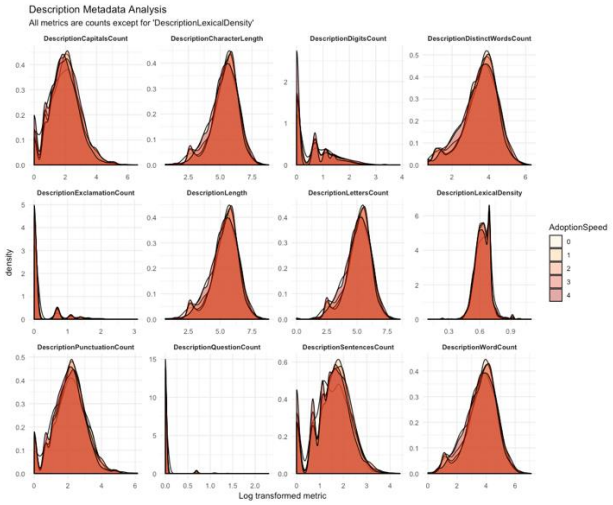

Figure 4. Metadata description analysis

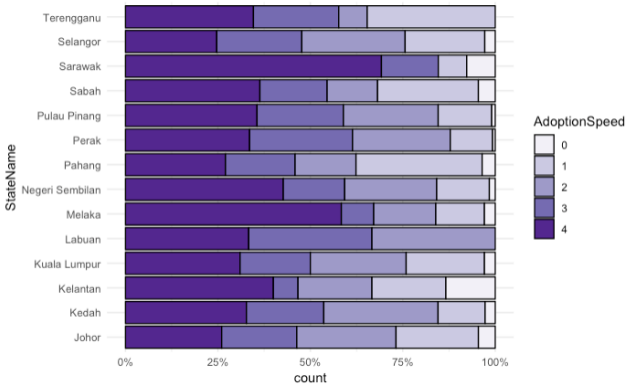

Figure 5. State percentages of adoption speed

V. Data Processing and Analysis

C. CSV Data: Natural Language Processing of the Description

Preprocessing: The description of each post is a sentence. The first step is to break down the sentence into words since the words with different capitalization are considered to be the same word. Therefore, we change everything to lowercase as well as remove all the stop words which are not meaningful to the study such as “the”, “on”, “is”, etc. Also, all the punctuations and extra white spaces are removed from the words list. Furthermore, we discount numerical values that show up in the description because we do not think it will be beneficial to the study. Lastly, we stem the words so that words like “love”, “loving”, and “loved” are all represented by “love.”

Term document matrix: Now, the description of each post becomes a list of words. Then, we create the term document matrix to represent this data. This is a way of representing the words in the description as a matrix of numbers. The rows of the matrix represent the posts to be analyzed, and the columns of the matrix represent the words from the description of the posts that are to be used in the analysis. The values in the matrix is binary, thus, 1 represents the presence of the word and 0 represents the absence of it. Therefore, in our matrix there are many zeros, and so we use a sparse matrix instead of dense. This is a more efficient representation of the information contained in the term document matrix. It is necessary for us to use this representation as there are a large number of words as it saves us computation time as well as memory space.

Rearrange adoption speed: Next, we reclassified the adoption speed group. If the adoption speed is between 0 to 3, we group it as 1, otherwise set to 0. In this new adoption speed group, the success (labeled as 1) means being adopted before 100 days of being listed, and 0 is the complement of that statement. The new adoption speed group variable is used as the response. The columns in the term document matrix is used as covariate. The idea is to fit a least absolute

shrinkage and selection operator, also known as Lasso regression using appearance of words in each post as predictors and whether the pet is adopted in 100 days or not as the response.

Lasso regression: Lasso regression adds penalty for non-zero coefficients. Since the response is binary, the actual model we are fitting is a logistic regression model, which is a type of a generalized linear modelization. It is different, however, from conventional multiple linear regression because we cannot use least squares method to fit the coefficients in the model. A traditional way is to find the fitted values for those coefficients through maximum likelihood. This maximization process is difficult, and therefore, it needs a maximization algorithm to complete is. Using Lasso, the way to fit the logistic regression model is similar. We still want to find the coefficients to maximize the likelihood function. However, we will add a constraint to the maximization problem. We require that the sum of absolute values of the coefficient has to be minimized. This type of problem optimization involves the Lagrange multiplier. Since Lasso penalizes the sum of the absolute values of the coefficients, the high values of this method will cause more coefficients to be exactly zero under Lasso. Therefore, this method is more useful for model selection compared to similar penalizing regression such as Ridge regression since other regression models will never be able to make the coefficients become exactly 0 even though the coefficients are insignificant. In other words, selecting the multiplier is important and actually decides how many coefficients we want to keep in the model. We use 10-fold cross-validation to select the best multiplier based on the mean-squared error. Once the Lasso regression is fitted, the predictors with the higher values in coefficients are considered to be useful to increase the chance that the pet is going to be adopted under 100 days. On the other hand, the lesser values are considered to be the words that lead the pets to take longer than 100 days to be adopted. The model can also be used to predict whether the pet is going to be adopted in 100 days or not; however, it can only predict the probability. Therefore, we need a cut off to tell how high the probability is that we can consider that the pet is going to be adopted in 100 days. We grid search the possible cut off from 0.1 to 0.9 with 0.05 increment and whichever gives us the lowest misclassification rate based on the training data will be the best cut off rate. The performance of the model is evaluated first based on the training data. The confusion matrix is provided to tell how good the model was. The final is applied to predict the test data.

Speed prediction: The goal is to use the description of the post to predict how quick an animal is adopted. We would like to see what kind of words are considered to be important for prediction. The top 10 most appearing words are shown in Table 1. It is not surprising that "adopted" is the most common word, and its frequency is much higher than that of "love" and 'home." The frequency table in Table 1 can further be shown more visually in the Wordcloud figure in Fig. 6.

Table 1
Frequency of Words in Description

adopt	11908
love	8209
home	8101

Figure 4. Frequency Wordcloud

dog	6359
cat	6167

The final model includes 718 non-zero coefficients.

From the result in the table, we can see that the mean square error (MSE) is the lowest when we include the term dog. Also, in order, the model will prefer to include dog first before kitten. Therefore, we can say that dog is more important to explain the variation in response, the rate of adoption. Also, when the multiplier increases, more predictors are going to be dropped or its coefficients become zero. The figure in the following shows that how does the change of the multiplier affect which coefficients or term will be included in the model. We can see, among the top 10 words, the word "pure" appears. It is not surprised that this is the first word to be included in the model. "pure" is most likely means whether the dog or cat is pure breed or not. People tends to have pure breed dog and cat since pure breed pets tend to have the traditional look, size and temperament. Also, pure breed pets do not have much health issues than mixed breed. In addition, it is surprising that we see the words like "email" and "whatsapp". It means that the way of communications is also important in order to have the pets got adopted or not.

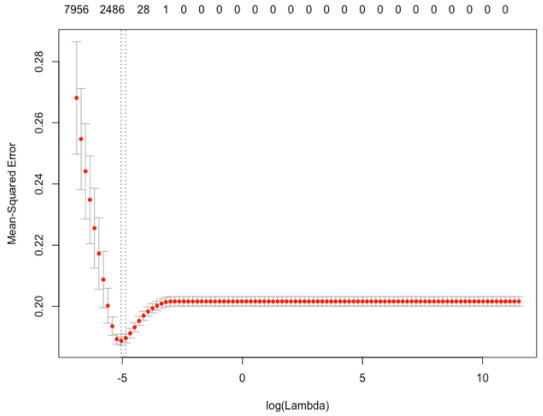

Figure 5. Lambda values with corresponding mean-squared error

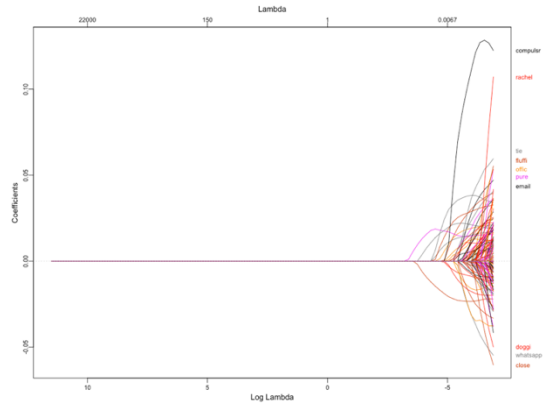

Figure 6. Lambda values and corresponding coefficients

Table 2
Models and Lambda Values

lambda	model
3511.192	dog
4229.243	dog, kitten
7390.722	dog, kitten, interest, puppi, must
8902.151	dog, kitten, interest, puppi, must, play
10722.67	dog, kitten, interest, puppi, must, play, found, cat
12915.49	dog, kitten, interest, puppi, must, play, found, cat, neuter, friend, compulsori

The final Lasso model with the selected multiplier is fitted. The terms of the highest 10 coefficients and the lowest 10 coefficients are shown in Table 3 and 4, respectively.

Table 3
Highest coefficient terms

term	coefficients
ago	0.683906
australia	0.153861
burmes	0.109048
darl	0.108556
built	0.091972
peanut	0.085627
desexualis	0.084007
mischeivi	0.076745
vaccin	0.072928
bobo	0.069124

Table 4
Lowest coefficient terms

term	coefficients
nelli	-0.42005
allah	-0.39012
pulainthiran	-0.37649
sebuah	-0.3435
puppiesbuddypixietwinkl	-0.33372
detailsmanythank	-0.32699
sheltermerci	-0.32093
pound	-0.29943
rahang	-0.28855
gst	-0.28694

Although they are not a complete vocabulary due to stemming process, we can still tell the meaning of few words. For example, “vaccin” is likely referring to vaccinated. That is certain an important factor people will consider when they adopt a pet or not. If they get a pet which is not vaccinated, then they still need to spend extra amount to get all the proper vaccines for the pets. Besides, some pets do not have the vaccine may actually due to other health issue such that they cannot get vaccines like other pets. Therefore, it is riskier for people to get a pet which is not vaccinated.

Another word that is interesting is “rascal”. “Rascal” is actually referring to a rascal trailer such as rascal bike pet trailer. It is a popular pet trailer for people to take the pets to outdoor when they bike. Some pets do not like trailer. If we look at the posts, we can find that some people post that “He likes to go with a ride in the rascal bike trailer”. Therefore, it could be a good way to advertise for the pet.

“Burmes” are actually referring to Burmese cats after we take hundred posts with the stem word “brumes” on it. Burmese cat is a breed of domestic cat. It is originating in Thailand. Burmese cats are extremely intelligent and alert, which means that they adore playing with toys and running around in playful mannerisms. This breed of cat, while very affectionate, does not mind telling its owners when it needs attention and will be very vocal about it. Therefore, it is reasonable to see this word “burmes” has impact on the adoption rate.

Now based on the description, we want to use this model to classify whether the pet is going to be adopted in 100 days or not. Since the model is going to predict the probability that the pet is adopted in 100 days. Therefore, we need to select a cut-off in order to classify whether the pet is adopted in 100 days or not. The idea is to grid search different potential choices of the cutoff values. The one provides the lowest mis-classification rate is the best cut-off.

According to the following figure, we see that the misclassification rate is decreasing when we change the cutoff from 0.1 to 0.65. After 0.65, the misclassification rate is dramatically increased. Therefore, the best cutoff is 0.65 with 22.76% misclassification rate.

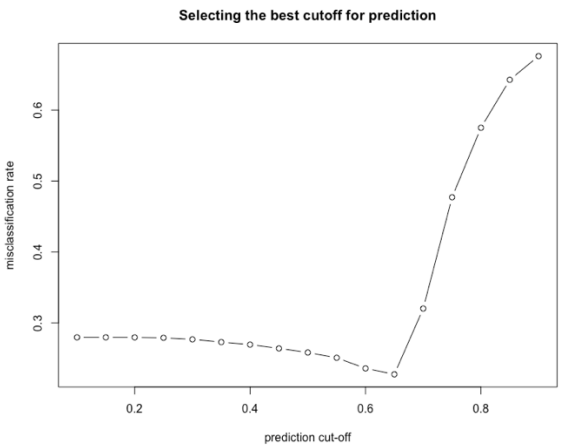

Figure 7. Cutoff rate for misclassification

Table 5
Training data confusion matrix

		True Label	
		adopted not in 100 days	adopted in 100 days
Predicted Label	adopted not in 100 days	1310	526
	adopted in 100 days	2887	10270

D. Image Data: Extracting Image Features

Introduction: In this part, we used a pre-trained neural network for the images. We have three major steps. First, we only take one picture of each adoption, which is the first picture of each PetID. Then, we pad the picture to square based on the features that Desnet121 and preprocessed input selected and ratio. Lastly, we resize our square picture to 256.

To speed up the process, we divided the pictures in small groups of 16 pages and after we finished our processing of one batch, we then moved onto another batch. We used TensorFlow pre-trained Densenet121 model to extract features from the images. Densenet121 model usually will output 1024 features after the GlobalAveragePooling2D, to simplify the result we then pool it again and took 4 features for each. Then, we saved our new 256 features (pixels) of each pet ID into a new data frame of both test pictures and train pictures.

Feature extraction algorithm to be used is first feed the image to a conventional pre-trained neural network (such as in TensorFlow), then use the representation for that particular image in the intermediate layers of the neural network.

Model selection: Then, with the image features data, we first apply outlier detection algorithms, SVM (one class SVM) and LOF (local outlier factor) and IF (isolated forest) to it. The number of outliers detected are as follows:

- SVM: 1459
- LOF: 1500
- IF: 1500

And the outliers and their indices in common of the above three algorithms are then extracted and deleted from the features data frame. The total number of outliers deleted is 438, which is about 2.9% of the feature data. This is reasonable to discard without losing much of the information for further analysis. After detecting the outliers, we apply normalization on the data for model selection. The next step is to apply cross validation which splits the normalized no outlier feature data into training and testing sets. Then, to apply machine learning models which

include regression models and other models to train image classifiers. It results in pairs of accuracy rates for both training set and testing set as seen in Table 6.

Table 6: Model selection training and testing results

Machine Learning Classifier	Training Set Accuracy Rate	Testing Set Accuracy Rate
Linear Regression	0	0
Polynomial Regression	No Result	No Result
Logistic Regression	0.8928203366540708	0.8639642734455514
LDA (Linear Discriminant Analysis)	0.4081071796633459	0.365166609412573
SVM (Support Vector Machine)	1	1
Random Forest Classifier	1	0.917554105118516
Decision Tree	1	1

From the results above, polynomial regression model does not work due to a very large running time required; other than that, linear regression model gives 0 accuracy rate for both sets which makes no successful fitting, and the reason might be that linear model fails when using raw image data from the beginning. Choices for neural network contains creating a feature vector from raw image, which provides a low-dimensional and noise-resistant way to represent images. Also, this dataset is used for classification, it makes sense that the regression model doesn't work. Thus, we will choose the best model from the other models which seem to work better. Among those working models, SVM, Decision Tree, and Random Forest Classifier returns very high accuracy rates, and they are of value 1 as well, which is interesting. With limited knowledge on Machine Learning modeling, we found that the possible reasons leading to the value 1 training set accuracy could be that there is some overfitting involved which is reasonable. However, the value 1 testing set accuracy are not looking nice, and the reasons could be that our image data used for modeling may be unbalanced so that the models can't function normally, for example, the image may not having the pet at most pixels but instead, have the adopter or shelter workers as main expression while the pet is just a little baby cat or dog. Another confusion finding from this accuracy rate chart that we so far can't explain why that is, Decision Tree Classifier has higher testing accuracy rate than Random Forest Classifier. Yet, so far, the Random Forest Model seems reasonable to use for prediction, and it could be chosen as the best model for the image data prediction.

VI. Conclusion

The goal of this project is to find things that correlated, preferable causing to the adoption speed of pets. With the effort on exploring all data types from the original data zip file, including metadata, csv data and image data, the original datasets are either used in direct analysis or transformed for deeper analysis. The interesting findings contain that some factors lead to faster adoption rate for pets. The metadata analysis confirms that cats are more likely to be adopted faster than dogs. The description data analysis suggests that “pure”, meaning purebred, pets are much healthier compared to mixed breeds and thus are faster to be adopted, while the human communication tools are not contributing much to the pet’s adoption process; also pets with longer description seem faster to be adopted. The csv data also shows that smaller, younger pets are adopted faster, as well as that vaccinated or dewormed situation does not matter much. The image data used with a TensorFlow pre-trained Densenet121 model for feature extraction isn’t showing significant findings on how the features affect the adoption speed, but outlier detection and model selection finally lead to a choice falling on Random Forest Classifier, meaning that further image analysis could consider using the Random Forest model as a prediction tool. There were hardship times when processing each type of data we encountered in this Kaggle data source, and successes and failures were both included along with analyses. For the reason that this pet adoption topic is interesting and widely containing different data types, there is a hope to continue it with our stronger analysis skills in the future.