Duke University the Fuqua School of Business
Data Science for Business
Predicting Adoption Speed – Pet Finder
October 13th, 2019

Team 38:

Yasi Chen, Shangyun Song, Lindsay Trinh,

Xinyi Zhu, Yuqian Hu, Jiali Yin

Business Understanding

Our group is currently acting as a consulting agency who works on behalf of PetFinder, a non-profit organization, contains a database of animals and aims to improve the animal welfare through collaborations with related parties. The core task of this project is to predict how long it will take for a pet to be adopted. According to PetFinder's 2018 Financial Report, approximately 70% of its total public support gained and revenues earned will be spent on its public welfare programs, including this animal adoption program. Our defined business problem is "When new pets come in, what would be the estimated time for new pets to be adopted." Based on different traits each pet has, we could estimate how long it takes for the pet to be adopted according to the model we created. Therefore, by adopting our model, the company can guide shelters and rescuers who post information about animals through the channel on estimating the adoption speed for new animals. Once the pet adoption speed is predicted well, more efficient resource allocation can be implemented to improve the overall adoption performance, and subsequently, reduce the costs of sheltering and fostering. Eventually PetFinder will receive more financial support and make a greater contribution to the global animal benefit.

Data Understanding & Preparation

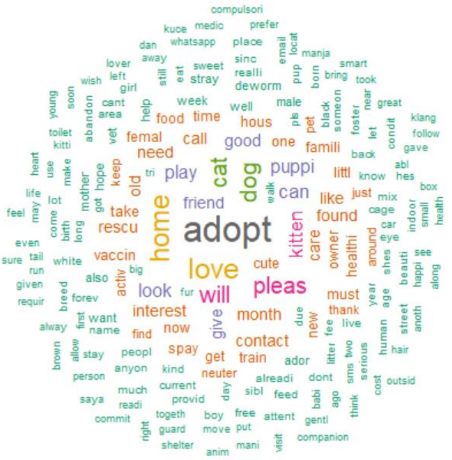
This dataset is retrieved from Kaggle, consisting of 14993 observations and 24 variables. All the data are based in the Malaysia area so our analysis is country-specific and culture-oriented. We removed the variable *Name* since we assume it is irrelevant to the pet's adoption speed. Other than *Name,* we removed *State, PetID,* and *RescuerID* for the same reason. In addition, some columns such as *Breed1* and *Type* are coded as numbers with an additional reference spreadsheet provided. For clarification reasons, we convert *Type* and *Gender* into

strings: we coded “1” as “Dog” and “2” as “Cat”; For *Gender*, we coded “1” as “Male”, “2” as “Female”, and “3” as “Mixed”. The mixed represents the gender of a profile of pets. We also changed categorical variables into dummy variables. Besides, we conduct a chi-squared test to test the correlation among *Color1*, *Color 2*, and *Color3*, the result shows that they are dependent on each other. Therefore, we plan to remain *Color1* and remove the other two. Finally, considering only 0.38% of records have *Breed2*, we remove *Breed2* for simplicity reasons. Besides, considering that the dataset contains a *Description* column, we applied text analytics to this column. Based on the analytics, we select the 10 most frequently occurring words and create an additional 10 columns based on these hot words. We check if each pet’s description contains these 10 words individually. If any word is matched in the description, we assign “1” under that column. We assign “0” if it does not. These 10 words are: *love*, *kitten*, *puppi*, *rescu*, *healthi*, *activ*, *cute*, *train*, *mother*, *kind*. Finally, we all agree that the length of a description may be an important factor as well. Since the more detailed the description is, the less time it takes adopters to pick the pet. Logically, adopters tend to make decisions quicker if the available information is richer. Therefore, our final dataset consists of 14,993 records and 29 variables.

Data Exploration

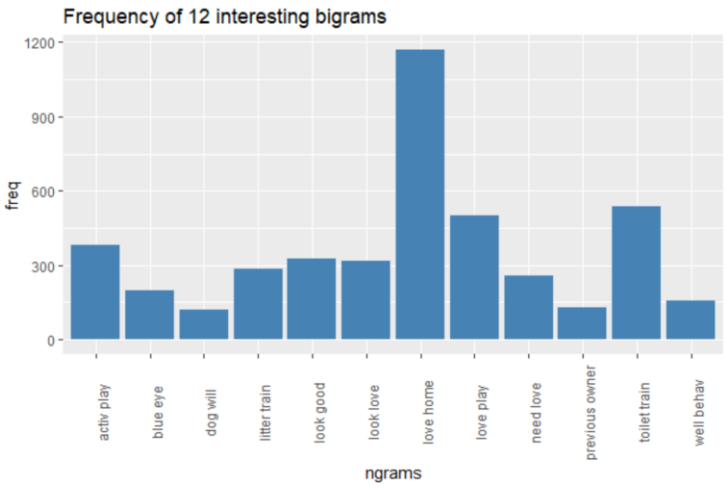
Text Analytics

Based on the text analytics, we first created a text cloud to capture the most frequent word in the *Description* column. It turns out that “adopt” appears most frequently, followed by words



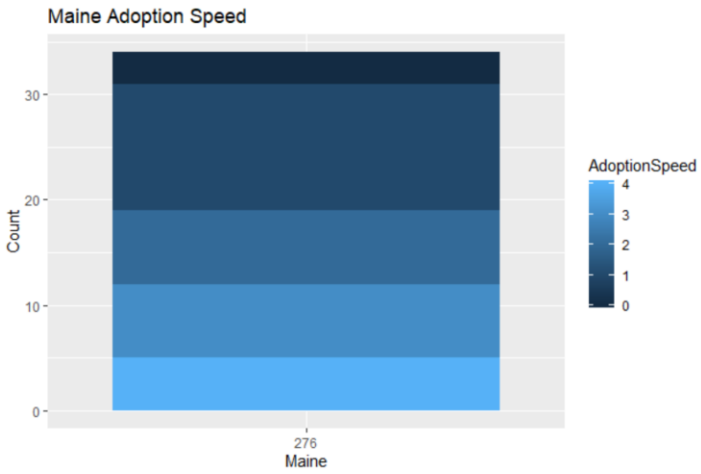
"kitten", "puppi" and "cute". From the word cloud, we can summarize the major characteristics of those pets in our dataset.

Next, we created bigrams to capture more patterns in the text. The most frequent bigram is "month old", and we pick 10 most intriguing words and plot them in a bar chart.



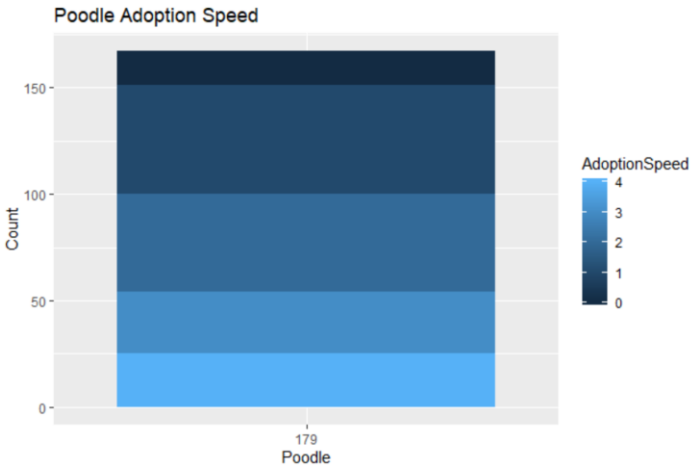
As we can see from the two graphs, rescuers and shelters tend to use a lot of descriptive words in the description such as *active*, *good*, *cute* etc. As mentioned above, to test the effectiveness of these words, we picked 10 hot words and converted them into 10 categorical variables. The reason that we did not choose bigram was that bigrams were made of stemming words, so it may be hard for the software to detect these bigrams in description.

Breed vs. AdoptionSpeed

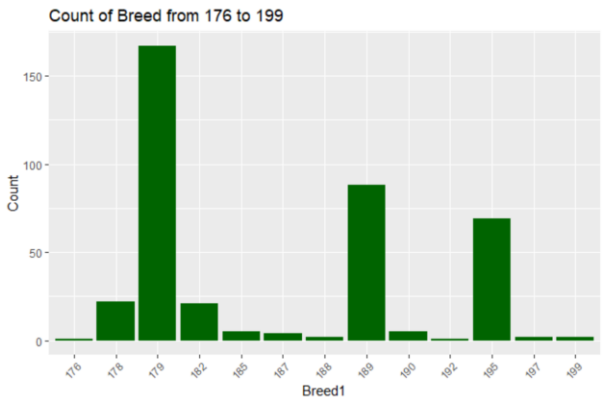


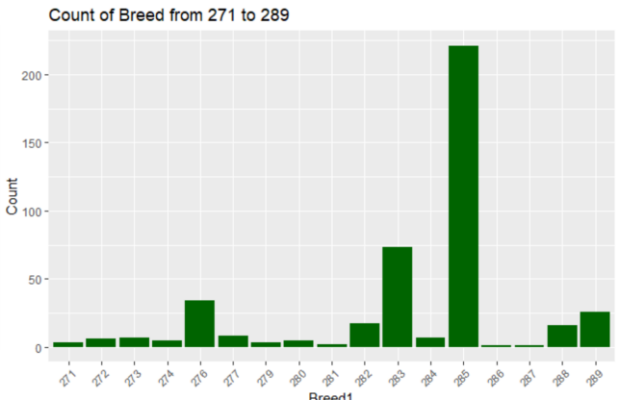
Additionally, from Perro Market Website[1], one of the most well-known pet websites in Malaysia, we learned the 10 most popular cat and dog breeds in the country. We picked Maine and Persian for Cat, and Poodle for Dog for the purpose of data exploration.

The mean of *AdoptionSpeed* for cats is 2.40 with a standard deviation of 1.21. By comparison, the average Adoption Speed for Maine is 1.97 and the average Adoption Speed for Persian is 1.95. Even though their mean is lower than the average, they are still in the first standard deviation range. Based on this fact, we could not conclude that popular cat breeds such as Maine and Persian tend to have shorter adoption period than other breeds.



On the other hand, dogs have an average AdoptionSpeed of 2.62 with a standard deviation of 1.14. Noticeably, Poodle has an average adoption speed of 1.97. Again, even though the average adoption speed is smaller than the average of the whole group, it still falls within 1 standard deviation range, so we could not conclude that popular dog breeds lead to shorter adoption period.



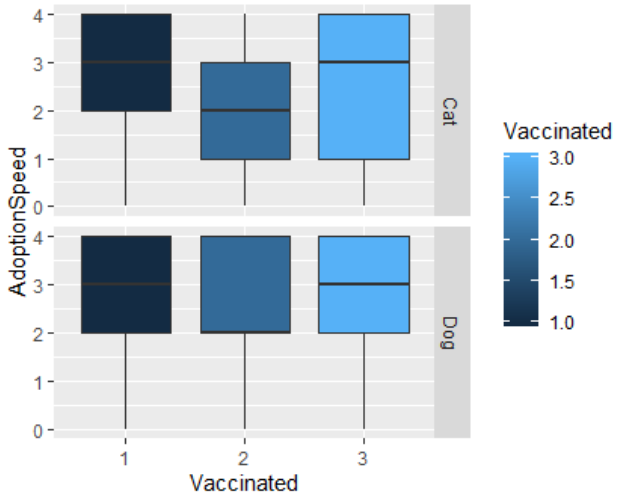


However, from the graph below, we find that the number of pets under popular breeds is much higher than the rest. Breed 179 is Poodle, and Breed 285 is Persian while Breed 276 is Maine.

Based on this preliminary analysis, our conclusion is that popular breed determines the number of pets under this breed but do not necessarily mean lower adoption speed.

Vaccinated vs. AdoptionSpeed

Vaccinated as "1" means the pet is vaccinated, "2" means the pet is not vaccinated, and "3" means not sure. The pet who has been vaccinated counterintuitively shows higher median of time span than the one not got vaccinated, and for those reported unknown they have higher time span waiting to be adopted. There's also no clear correlation (-0.06) between the *AdoptionSpeed* and *Vaccinated* across the species.



Modeling

Since the target variable *AdoptionSpeed* is a categorical variable with 4 levels, we decided the core task of this project to be **classification:** when a new pet comes in, the model would determine how long it will be adopted based on which level of adoption speed it is classified into. Because of its multi-class characteristic, we choose three models to achieve the goal, including Random Forest, SVM, and Multinomial logistic regression. All three models help solve our business performance by predicting the adoption speed for new animals based on their provided attributes. By sharing our predictions with the company, the company can have a better

plan about the allocation of current resources, especially allocating more resources to the animals with longer predicted adoption period to accelerate their adoption speed.

Random Forest

Random Forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. One pro with Random Forest is that its decorrelation when dealing with highly correlated variables. One con is its limitation on the number of categorical variables could be included in the predictors. In our project, we excluded the *Description* since there are more than 14000 levels in one variable and it has already segmented into words with high frequency and description length.

Support Vector Machines (SVM)

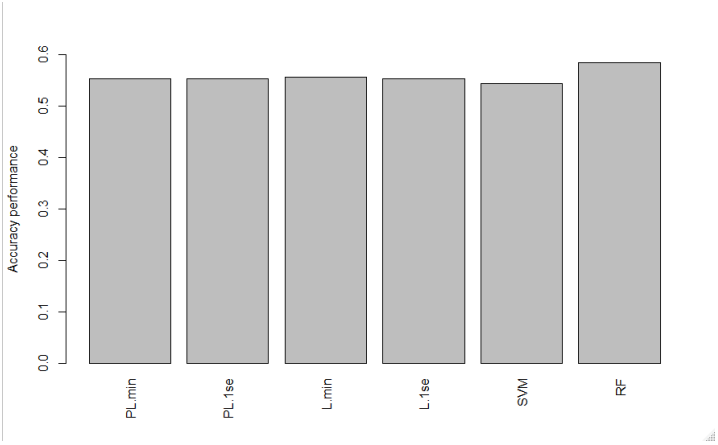
Another model we used is the Support Vector Machines (SVM) model, to create the optimal hyperplane to separate classes and classify new records in our test dataset. We include all the variables from the cleaned training dataset, except *Description*. SVM is easily adopted in solving multiclass classification problems by transforming complex data into high dimensional analysis and thus having classes easily separable. The model itself has good computational properties, but it does not attempt to model probability. Comparing with Random Forest, SVM can only be used to predict the class instead of estimating the probability of different classes. From this point of view, Random Forest may offer us a more comprehensive understanding of the prediction of classes of the test dataset.

Multinomial Logistic Regression with Lasso and Post Lasso

We also used multinomial logistic regression to predict our target variable. However, because of the limitations on the number of variables could be included in multinomial logistic

regression, we used Lasso to select related variables for each class of adoption speed. We built 4 models of multinomial logistic regression with Lasso and post-Lasso based on minimum λ and 1 standard error λ. After selecting the variables, the models eliminated overfitting problems and may make better predictions. However, the variables selected by Lasso may differ in different bootstrapped models, and they are hard to be interpreted.

Evaluation



In order to evaluate the performance of each model, we conducted a 3-fold cross-validation and selected the Confusion Matrix Balanced Accuracy as our Out of Sample accuracy measurement. Based on the chart above, all of our six models have a similar level of accuracy performance, ranging from 0.5 to 0.6. Random forest achieved the highest level of OOS accuracy of 59%. Multinomial Logistic Regression with Lasso using minimum λ is the second highest with 56% OOS accuracy. Considering OOS accuracy as an important indicator for the classification models, we chose Random Forest as the optimal model to predict a newly coming pet's adoption speed. To investigate specific classification accuracy, we further evaluated the confusion matrix of Random Forest and SVM, one has the highest and one has the lowest OOS accuracy.

Random Forest

Prediction \ Reference	0	1	2	3	4
0	3	0	1	1	1
1	50	318	282	157	125
2	29	419	522	356	290
3	13	98	207	231	125
4	36	193	322	316	902

SVM

Prediction \ Reference	0	1	2	3	4
0	0	0	0	0	0
1	30	167	119	91	62
2	70	687	891	617	688
3	8	28	55	90	53
4	23	146	269	263	640

From the confusion matrix, we can conclude that Random Forest predicts more accurately when pets are actually adopted within 1 month (AdoptionSpeed = 2) and no adoption after 100 days being listed (AdoptionSpeed = 4). Random Forest does not show a tendency towards one specific class. In contrast, SVM is leaning toward classifying pets into AdoptionSpeed level of 2.

One possible limitation with the model is that there are too few data points with 0 *AdoptionSpeed* and therefore it's hard to evaluate its classification accuracy. On the other hand, with the above-mentioned limitation of Random Forest, if the PetFinder adds a new categorical variable that has more than 100 levels, Multinomial Logistic Regression with Lasso using minimum λ, instead of Random Forest, would be a better choice to classify newly coming pet's adoption speed.

When we try to apply our model to real-life business, we should develop a business case to identify if the prediction is accurate and if the business can utilize the prediction to optimize its resource planning and thus improve the adoption performance. The business can compare its previous adoption performance with its future adoption performance by having resources allocated based on our prediction model. The improvement in adoption performance can be proved by the increase in the number of animals adopted within the same period of time. In

addition, the business should analyze if the animals with long predicted adoption time can be adopted within a shorter period of time, to see if the change in resource allocation is efficient.

Deployment

By evaluating each model with its OOS accuracy, we reached a conclusion that Random Forest, which generated the highest OOS accuracy, would be the optimal choice for PetFinder. Whenever a new pet comes in, PetFinder is capable to predict how long it will take for a pet to be adopted by simply inputting its features. Estimating the adoption time more accurately will enable us to improve the animal's profile, which is crucial to expedite the adoption speed, and subsequently save the costs of sheltering.

In order to have more insights on what specific features have a significant impact on pets that take a longer time to be adopted, we examined the variables selected by Multinomial Logistic Regression with Lasso using minimum λ. Since the variable selection differs in each fold of data, we took the intersection of the variable selections in 3 folds and found that across the four levels of adoption speed, *Age*, *Breed*, *Amount of photos uploaded*, and *Amount of Video Uploaded* are important. Specifically, for pets with slower adoption speed, people pay more attention to their health: whether they are vaccinated, dewormed, and sterilized. Also, the length of the description plays a more important role in the profile of pets with slower adoption speed. If pets are classified as possibly taking a longer time to be adopted, PetFinder should encourage the shelters to prioritize the health of the pets and increase the length of the description on profile.

One important ethical consideration is that if the pets should be treated differently based on their predicted adoption speed. Without the prediction model, we assume that all pets are

receiving an equal level of treatment. However, if we suggest the business to allocate more resources to animals with lower predicted adoption speed, we are resulting in an unequal distribution of resources. If our prediction is not accurate, some pets will be negatively influenced and more unlikely to be adopted. From this perspective, the business' goal of improving global animal benefits may not be fully achieved.

One possible risk for our model is that we are currently having an inaccurate prediction for pets in the actual adoption speed class of 0. If those pets have predicted class of 3 or 4, the business will allocate more resources to promote them and thus results in a waste of limited resources. To alleviate the risk, we should always update and improve our model as we have more records, especially records with 0 actual adoption speed.

Appendix:

3 Folds Accuracy Performance

	PL. min	PL. 1se	L.min	L.1se	SVM	Random Forest
1	0.5558069	0.5558069	0.5579374	0.5552764	0.5418801	0.5896837
2	0.5535119	0.5535119	0.5576338	0.5520558	0.5395327	0.5869292
3	0.5529835	0.5529835	0.5546449	0.5534739	0.5485869	0.5755870

Reference:

[1]: https://perromart.com.my/blogs/perro-learning-center/popular-cat-breeds-in-malaysia