Pet Adoption Metadata Analysis

Jue Wang, M.S.

Nima Zahadat, Ph.D.

The George Washington University

**ABSTRACT**

The purpose of this project is to demonstrate data mining techniques, visualization, and Natural Language Processing (NLP) using the metadata of pet adoption from the Kaggle website for pet adoption speed prediction and trends. Visualization focused on the following data: age scaled in months, cat or dog, colors of pet, gender, maturity size, fur length, whether vaccinated or dewormed or sterilized, any injury, top 5 breeds of cat and dog separately, name, and adoption speed. NLP Modeling focused on the following data: description of pets and adoption speed. Multinomial Naïve Bayes Classifier correctly assessed the adoption speed based on descriptions of pets with 35% accuracy. The ensemble method to find the best Random Forest model with 2-Grams vectorizer and TF-IDF vectorizer also applied. The Random Forest Classifier predicted adoption speed based on descriptions with 45% of accuracy. Finally, the limitation and further improvements are provided, and the review of results is extracted.

Keywords: data mining, visualization, NLP, Random Forest Classifier, Multinomial Naïve Bayes, pet adoption

**INTRODUCTION**

According to the American Society for the Prevention of Cruelty to Animals (ASPCA), around 4 million dogs enter shelters every year (ASPCA, n.d.). Many companion animal seekers acquire their new family members from shelters (Weiss, Miller, Mohan-Gibbons, & Vela, 2012). The criteria applied for holding a dog in a shelter for adoption are comprehensively subjective and will consider the characteristics such as breed, age, gender, size, temperament, health, and appearance (Winograd, 2009). Meanwhile, Pet Food Manufacturers Association (2012) reported that almost 47% of pet owners will relinquishment the ownership of their dogs and turn them to the agents like breeders, pet shops, who can resell puppies or dogs. Some owners explain that with finding a new job, they need to relocate to other states, with different apartments, which does not allow pets.

Without any adoption happened to those pets, they will be crucially euthanized due to the limited space and any other food or financial resources in shelter or any pet adoption organization. Or for other reasons, such as no-kill shelters, non-aggressive breeds, or having a higher likelihood to be adopted pets, shelters will keep those pets until they are adopted (American Veterinary Medical Association, 2002). However, the criteria applied for holding a dog in a shelter for adoption are comprehensively subjective and will consider the characteristics such as breed, age, gender, size, temperament, health, and appearance (Winograd, 2009).

In this article, I will evaluate the animal adoption speed from the Kaggle website. The effects of the physical characteristics of age, gender, breed, color, fur length, maturity size, health condition, whether the pets are being vaccinated, dewormed or sterilized and using machine learning tools to analyze associations between description and adoption speed.

**LITERATURE REVIEW**

There are numbers of research articles related to the relinquishment of dogs. Some dogs are abandoned because of their behavior problem towards children or other companion animals. Those behavioral issues are including but not limited to paying attention toward owner behavior, barking or aggressive reaction toward other pets. Coe et al. (2014) reviews earlier relinquishment literature from 13 shelters in the United States and summarized that relinquishment of dog and adoption failure occur are significantly related to the dog’s behavior, particularly those related to other companion animals and children. Salman et al. (1998) complement this by examine another 12 shelters in the United States and notice that the lack of obedience training from the previous owner is the main reason that those pets would not be adopted or abandoned. Later, Shore (2005) studied 78 fail-to-adopt adopters in the Midwestern area of United States, the main reason to take a back seat to adopt pets is due to the behavioral problem related to children and other companion animals.

When adopter tries to successfully adopt a pet, they also take into account the pet’s characteristics. Previous studies explain factors that could affect dog adoptions. At the early stage, Posage, Bartlett, and Thomas (1998) exposed that some features of animals that appeared to affect the adopter’s decision to adopt or not included the size of the dog, which shows smaller dogs having a higher possibility of being adopted rather than larger or extra-large dogs; the color of the dog, which indicates the white, grey, or gold color dogs have a higher likelihood of being adopted than any other colors; and the breed of the dog, which discloses that toy, terrier, and non-sporting breeds are more possible to be adopted by adopters. Age will be one of the major considerations when adopters are going to adopt a dog. Hart, Takayanagi, and Yamaguchi demonstrate that age of dog appears to attract adopter’s attention from other characteristics before adoption and puppies are preferred to adult dogs. Later Diesel, Smith, and Pfeiffer (2007) investigate the length of stay in the shelter and time to adoption of dogs. The factors are similar to the previous researches: length of stay in the shelter for every dog is involved by dog breed, size, gender, age, purebred or mixed breed, whether a dog is neutered or spayed or not. In work by Siettou, Fraser, and Fraser (2014) indicate that factors like dog size, age, fur length, background, highly influence the results of whether a pet will be adopted or not.

Even though the cat will not have such intensive behavior issue the dog has, the reason of failure adoption of a cat exists also in terms of temperament, personality, and appearance. A cat with playful, approachable, younger, lighter-colored male and purebred characteristics has a higher possibility of adoption than a cat with not playful or older, darker-colored female characteristic (Weiss et al., 2012). Brown and Morgan (2015) examine the cat’s length of stay in shelter, and conclude that kitten will have a greater likelihood of adoption than juveniles; male cats will be adopted faster than female cats, and neutered males will even faster than male cats to be adopted; ‘Exotic’ breeds such as Persian, Russian Blue, and Ragdoll and Siamese cat will stay shorter in shelter; light-colored cats are more likely to find a new home than black cats.

There are a few words that periodically appear in the literature as potential factors to fail in adoption. Contributors to particular interests include age, and maturity size, breed, color. By incorporating these two and other variables into analysis.

**RESEARCH METHODOLOGY**

Here are the basic research methods I applied to the dataset:

* Data cleaning and data mining
* Visualization on variables
* Natural Language Processing (NLP) with modeling

Python 3 and packages of pandas, numpy are used to clean the raw train and test CSV files and combine two data frames into one data frame by defining ‘DataType’ variable as train or test. Data cleaning process are used to replace the integer values for some feature variables, to rename null values to some appropriate Python data type of string. And then, by applying data mining process, only the useful variables for both visualization and NLP are remained in the cleaned data frame and save it as a CSV file for future usage (for NLP, another data cleaning is needed). Visualization of all variables is created by Tableau, Python packages of matplotlib and seaborn, plotly and d3. Visualization provides the basic Exploratory Data Analysis - the fundamental patterns of the dataset and offers core information among variable relationships. Next, applying NLP is done using Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSI) models on the specific variables from the cleaned dataset that are not used in visualization (‘Description’ variable) in order to understand the text by analyzing topics. Then, apply the Multinomial Naïve Bayes Classification to fit and predict the model. Last, another NLP modeling is done by using the ensemble method of machine learning to select the best Random Forest model on the data for multi-label classification. In the end, model analysis and conclusion are drawn based on the previously mentioned processes.

**DATA**

PetFinder.my Adoption Prediction is acquired from Kaggle.com (PetFinder.my, 2019). The original row datasets contain 25 features and 14,993 samples in train.csv and 24 features and 3,972 samples in test.csv. Through the data cleaning process, the cleaned datasets contain 21 features and 14,993 samples in train.csv and 20 features and 3,972 samples in test.csv. In cleaned train dataset, the features are ‘AdoptionSpeed’, ‘Age’, ‘Breed1’, ‘Breed2’, ‘Gender’, ‘Color1’, ‘Color2’, ‘Color3’, ‘MaturitySize’, ‘Fur Length’, ‘Vaccinated’, ‘Dewormed’, ‘Sterilized’, ‘Health’, ‘Quantity’, ‘Fee’, ‘Description’, ‘PetID’, ‘DataType’, and there is no ‘AdoptionSpeed’ variable in test dataset. In the original dataset instructions, breed\_labels.csv and color\_labels.csv list the number and corresponding breed name or color, so the resulting cleaned dataset removes all number in ‘Breed1’, ‘Breed2’, and 3 color-related columns, and insert the top 5 breed name and all colors correspondingly. The resulting cleaned datasets focus on the top 5 breeds for cats and top 5 breeds for dogs in the ‘Breed1’ variable and change the related integer values in ‘Breed2’ variables, due to the tremendously large amount of breed listed and involved. The top 5 breeds for cat or dog samples cover at least 75% of both test and train datasets together, and many of the other breeds have less than 5 samples per breed name, or even 1 sample per breed name. Besides, variables ‘Type’, ‘MaturitySize’, ‘FurLength’, ‘Vaccinated’, ‘Dewormed’, ‘Sterilized’, and ‘Health’ are all listed as integer from 0 to 3 or 4, so the resulting cleaned datasets also replace the integers with matching values in Python data type of string. A list of features that was applied in this project, associated with the type of data, and a variable description were available in Appendix A. The detailed descriptions of prediction variable ‘AdoptionSpeed’ are listed in Appendix B, noticing that there is not pet data available for adoption speed between 90 and 100 days. The dataset organization’s overview indicates the original intent is to predict the adoption speed of a pet - how fast will a pet be adopted by applying the developed algorithm (PetFinder.my, 2019)?

**DATA ANALYSIS**

The variables ‘PhotoAmt’, ‘RescuerID’, ‘State’, ‘VideoAmt’ are four features not being used to do visualization or NLP and are dropped from the cleaned dataset. All variables except ‘AdoptionSpeed’, ‘Age’, and ‘Fee’ have input values as string. ‘Name’ variable is organized uniformly by condensing values of ‘No name yet’, ‘No Name Yet’, ‘no name’ and NaN into the same ‘No Name’. Visualizations including line chart and pie chart are created by Plotly wrote in Python. Correlation, scatterplots and word clouds are generated by seaborn and matplotlib packages in Python. Tableau is used to create some interactive plots for every categorical variable and d3 is used to generate another interactive histogram and pie chart. The NLTK package in Python provides support for all data cleaning process on the ‘Description’ variable for NLP, including punctuation removal, tokenization, stop words removal, lower cases, stemming and lemmatization. First, a Multinomial Naïve Bayes Classification was applied to predict and fit variables to classify ‘AdoptionSpeed’ variable across split train and test. The sklearn Python package provides support for majority parts of modeling requirements, such as splitting data into train and test, vectorizing n-gram and Term Frequency Inverse Document Frequency (TF-IDF), find best evaluation model, fitting split train data, and plotting confusion matrix. Last, ensemble methods with 2-gram vectorizer and TF-IDF vectorizer are deployed to find the best parameters for Random Forest model, then do the fitting model, predicting model, plotting the confusion matrix and check two best models’ differences. The code is available on Github: <https://github.com/juew72/pet-adoptionspeed>

**KEY FINDINGS**

Visualization on variables provides multi-dimensional information to obtain some basic ideas of the relationship between categorical variables and ‘AdoptionSpeed’ variable. The correlational plot on all variables is created to assess the positive or negative relations among variables. Figure1 demonstrates the importance of each variable on inflecting ‘AdoptionSpeed’ and the following information can be obtained: ‘Type’, ‘FurLength’, ‘Vaccinated’, ‘Dewormed’, ‘Sterilized’, ‘Fee’ have negative correlation with ‘AdoptionSpeed’ and ‘Age’, ‘Gender’, ‘MaturitySize’, ‘Health’ have positive correlation with ‘AdoptionSpeed’. As mentioned in Data section, 2-breed and 2-color columns, even though the row data has values as the number in these five columns, it only means the breed or color name, so remove breed and color variables for correlation consideration. Among those variables, ‘Type’, ‘Age’, and ‘FurLength’ have relatively strong correlations with ‘AdoptionSpeed’ compared to other categorical variables, with ‘Age’ having the highest correlation coefficient as 0.1, indicating the ‘AdoptionSpeed’ value increases as the age of pet increases; adding a month to pet’s age has a significantly positive, correlational effect on the other, which leads to increase the value of ‘AdoptionSpeed’ and make the pet wait longer to be adopted. It is no surprise that ‘Health’ and ‘MaturitySize’ have positive correlations with ‘AdoptionSpeed’, though the correlation coefficient might be small. People are willing to adopt a pet without any injury happened to it and also prefer to have a small size pet rather than an extra-large pet.

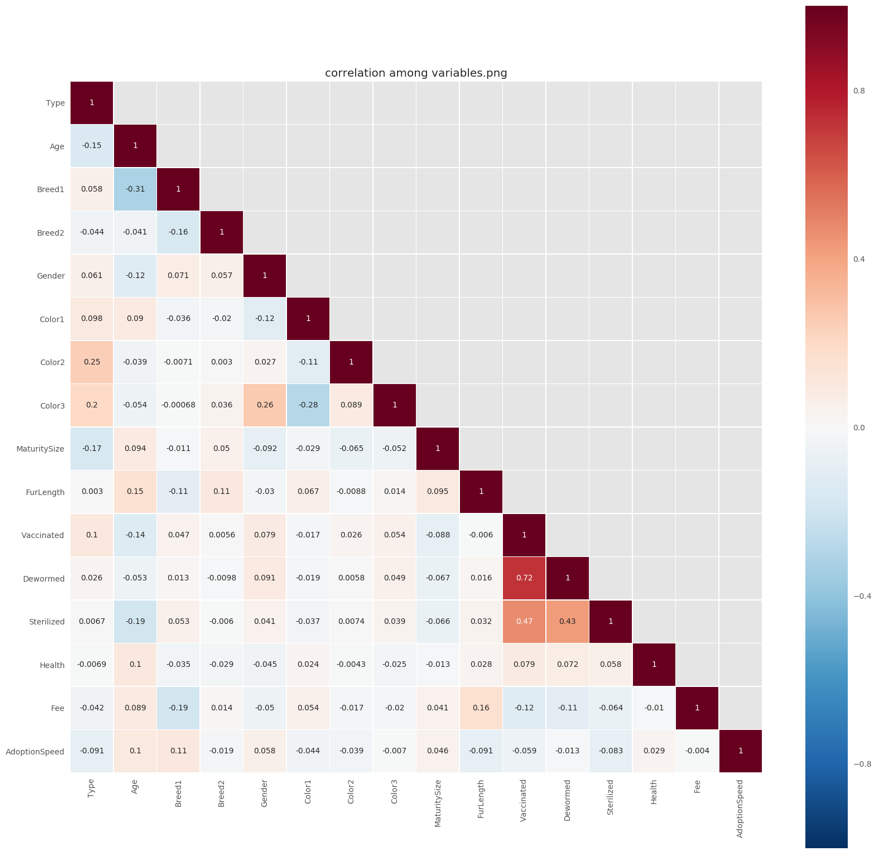


Figure 1. Correlation among variables

Next, pie charts, histograms, scatter plots, and line charts are useful in determining the nature of the interaction between the ‘AdoptionSpeed’ variable and each of the categorical variables under consideration. ‘Type’ variable is used as the stratification value between ‘AdoptionSpeed’ and the other categorical variables and also applying analysis on ‘AdoptionSpeed’ variable:

* ‘AdoptionSpeed’

Figure 2 indicates some underlying features for ‘AdoptionSpeed’ variable. For both dog and cat, the same day adoption has 2.73%; adopted within the 1st week has 20.6%; adopted within the 1st month has 21.7%; adopted within 2-3 months has 26.9% and not being adopted has the highest 28%. Some pets are adopted immediately but this is a rare case. The reason it happened might because of the eager to adopt from people. Someone is willing to adopt a dog or cat, or listed pet is exactly what he or she wants from every aspect, such as breed, color, size, fee, etc.

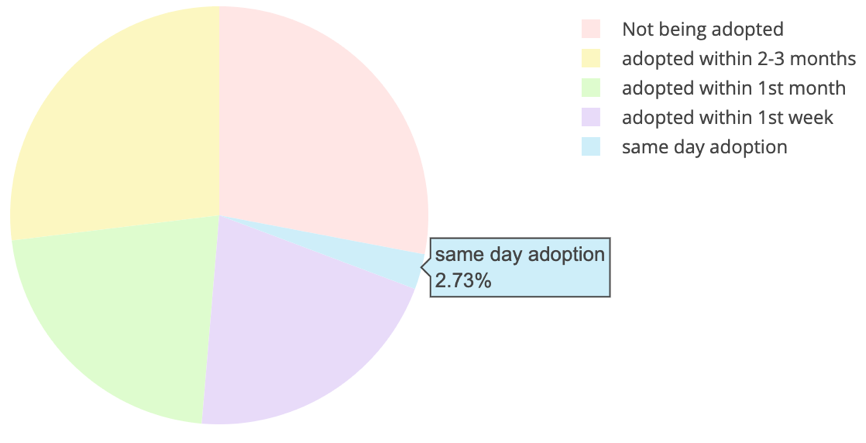


Figure 2: Adoption Speed

Figure 3 below shows cat and dog adoption speed for various adoption speed. It shows cat has higher adoption speed than the dog on the same day adoption and adoption happened in the first month. The total number of cats not being adopted is much smaller than the total number of dogs not being adopted (1783 vs 2414). This might because of people’s conception of human-companion animal relationship, and the everyday efforts that required in keeping a good companion-animal guardianship (O’Connor, Coe, Niel, & Jones-Bitton, 2016). The research shows the most of dog adopters assume more effort required and less easiness to take care of pets than cat adopters believe.

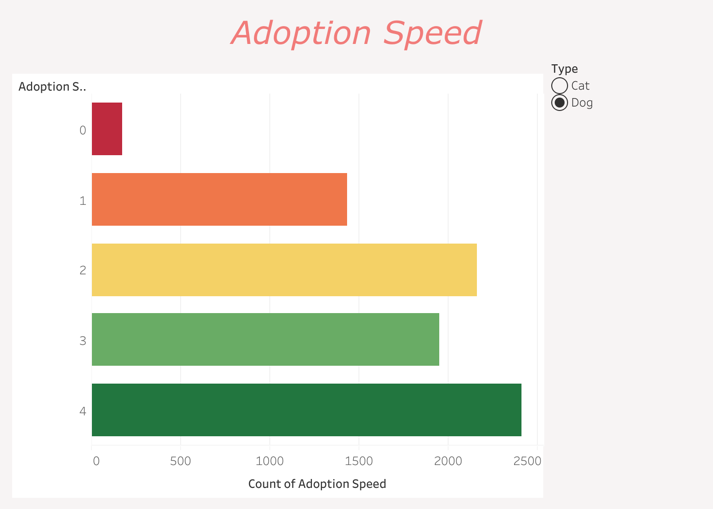
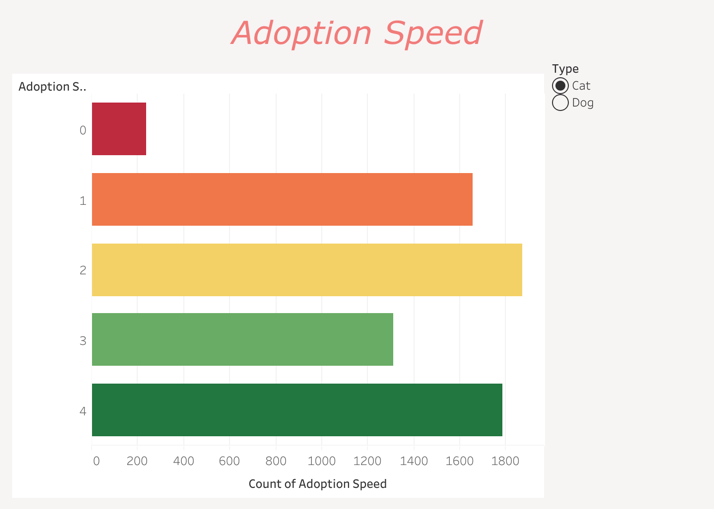


Figure 3: Cat/Dog vs Adoption Speed

* ‘Age’ and ‘AdoptionSpeed’

Figure 4 below shows the fundamental information about age for cat and dog. Most of the pets, no matter cat or dog, are young when they are listed, and most of them are less than 25 months old, which is around 2 years old. Even though it is an outlier for both dog and cat, the total number of pets will reach a peak when pets are aged in 2 months.

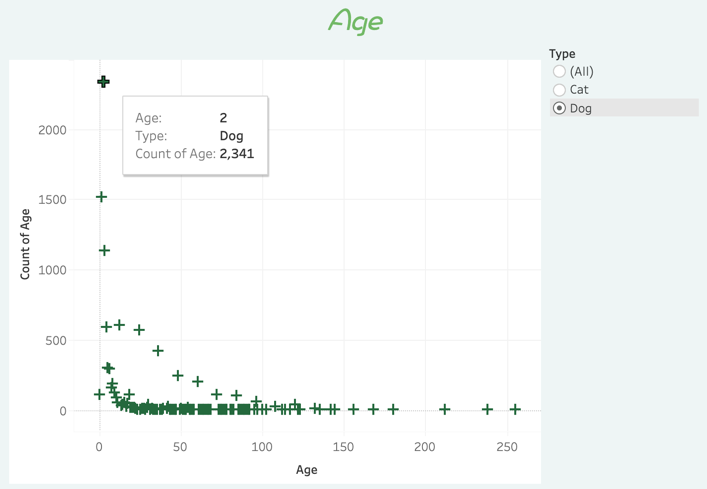
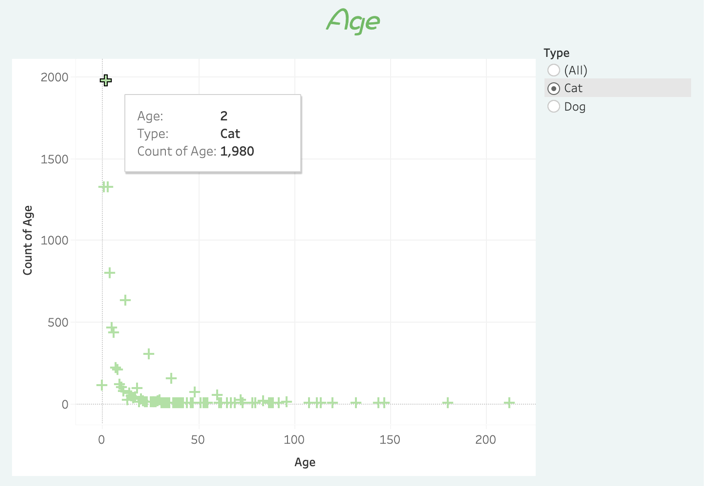


Figure 4: Cat/Dog age

Figure 5 listed below shows the relationship between age and adoption speed. Most of the pets are adopted within the first 60-month-old, which equals to 5 years old or younger. These young pets are adopted fast and most of them are adopted. There is a huge adoption at the age in 5-month, and they are adopted within 2-3 month, which is consistent with the results of some other studies (Brown et al., 2013; DeLeeuw, 2010). (Diesel et al., 2007) suggests puppies have better looking and are more likely to appeal adopters, that is why they are adopted much faster than older pets. So, there does appear to have a weak relationship between the pet age and the adoption speed, with younger pets, both cat and dog, having a higher possibility of being adopted.

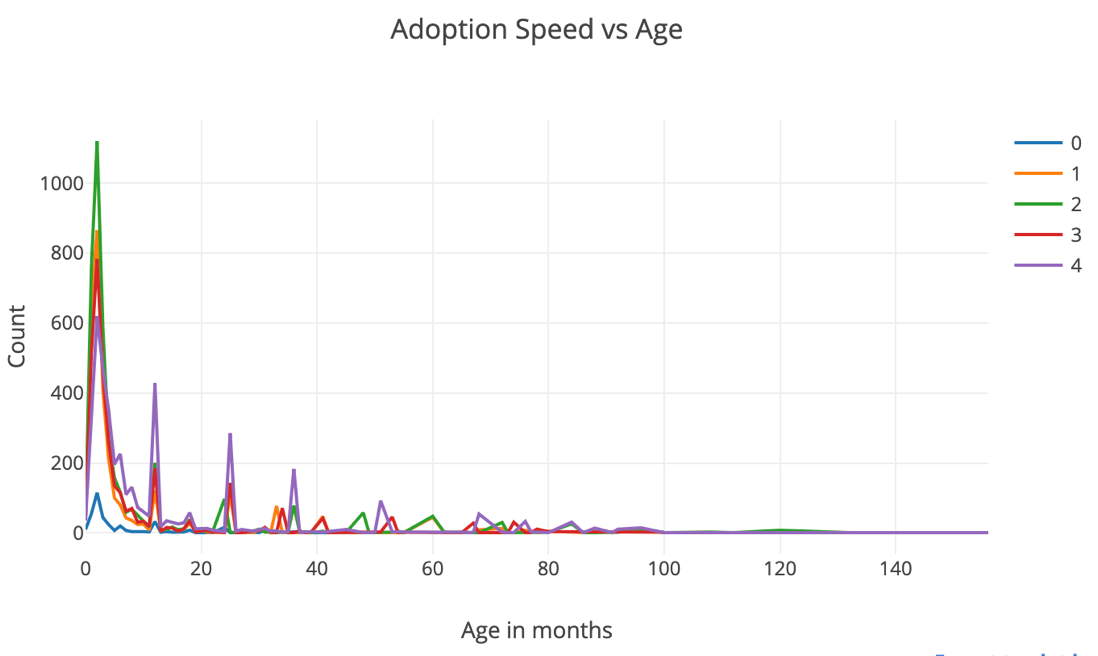


Figure 5: Age and Adoption Speed

* ‘Gender’ and ‘AdoptionSpeed’

Figure 6 below shows the relation between gender and adoption speed. For the cat, there are 4008 female, 3268 male and 1659 neutered or spayed; for the dog, there are 5137 female, 3773 male and 1132 neutered or spayed. In total, the female and male pets are mostly adopted within 1-to-3-months, but male pets are adopted faster than female pets. (Lepper, Kass & Hart, 2002) states that the adopter prefers a female cat or dog over male pets. (Žák et al., 2015) also indicates that male dogs remain significantly longer compared to abandoned female dogs. And there are more female pets than male pets, it is reasonable that more female pets are adopted in each stage of adoption speed.

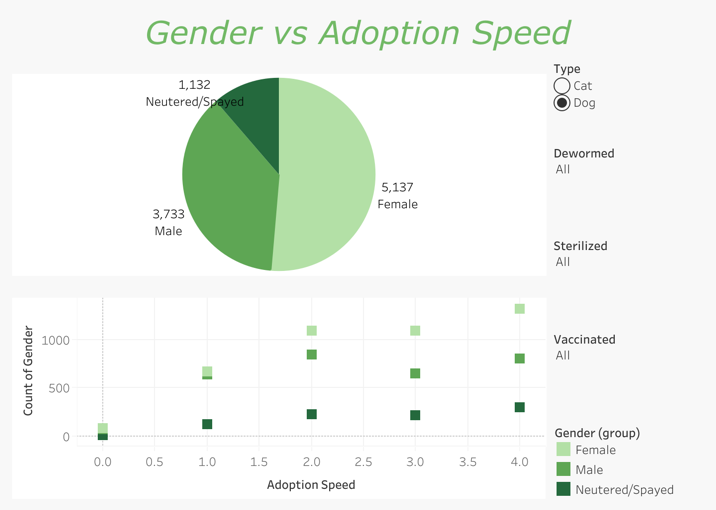
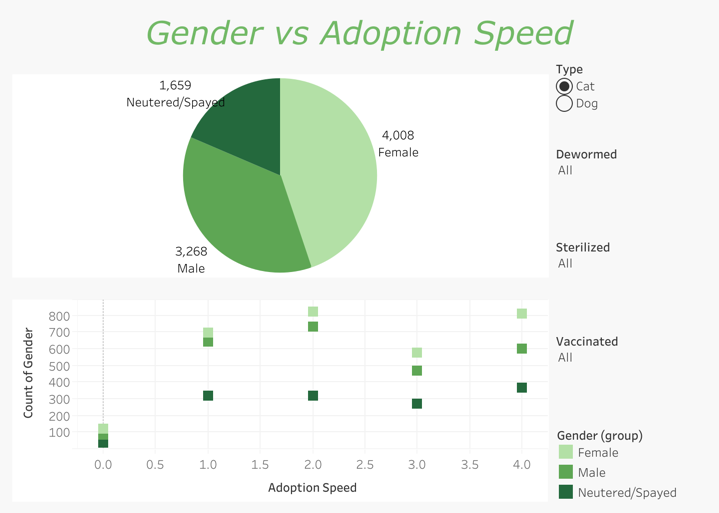


Figure 6: Gender and Adoption Speed

* ‘FurLength’ and ‘AdoptionSpeed’:

Figure 7 below shows the relation between fur length and adoption speed. There are more short hair pets than medium or long hair pets. For short hair pets, the largest amount of them is adopted within the 1-to-3-month. However, pets with long hair tend to have a higher opportunity to be adopted, which is consistent with the correlation: the ‘AdoptionSpeed’ value increases as the fur length of pet increases; for example, adding some lengths from short hair to medium hair has a slightly negative, correlational effect on the adoption speed, which makes the pet wait shorter to be adopted.

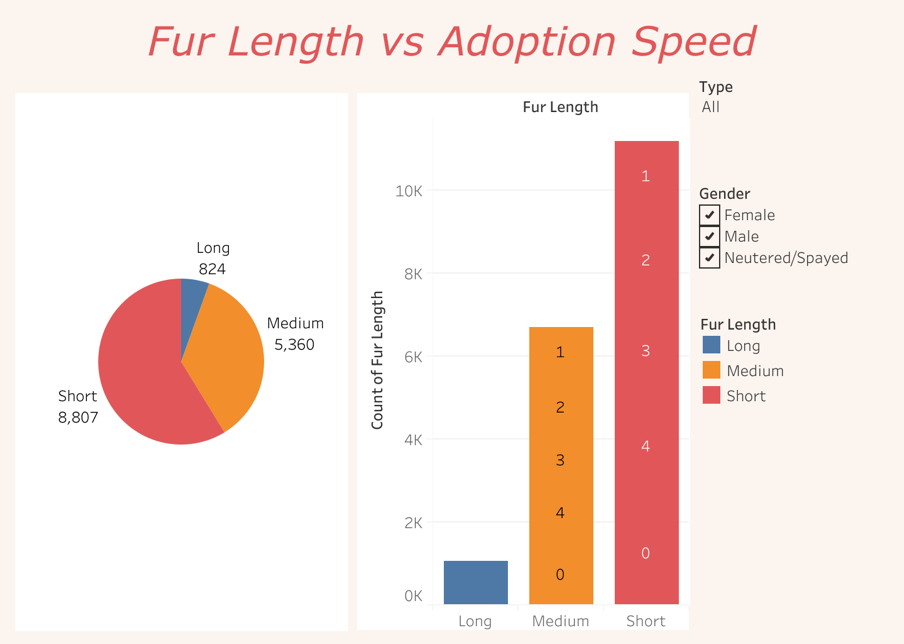


Figure 7: Fur Length and Adoption Speed

* ‘MaturitySize’ and ‘AdoptionSpeed’

Figure 7 below shows the relationship between maturity size and adoption speed. The medium size is the most common maturity size among all four sizes for pets. And there are almost no Extra-Large pets. The adoption speed of medium size pets has slightly higher than the others. Since there are 2973 not being adopted medium size pets, which is the largest number in medium size. For both medium and small size pets, it is more likely to be adopted within the 1st month, which is consistent with the results of some other studies (Žák et al., 2015): emphasizes that small dogs and medium dogs have the shortest length of stay and large dogs has the longest length of stay.

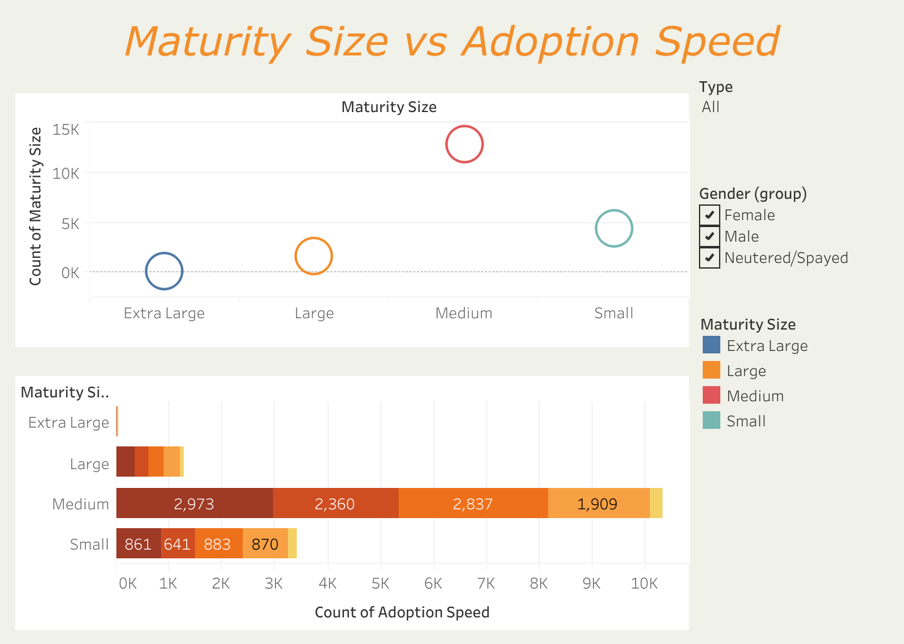


Figure 8: Maturity Size and Adoption Speed

* ‘Breed’ and ‘AdoptionSpeed’

Figure 7 below shows the relation between breed and adoption speed. Non-pure pets have higher and faster tendency to be adopted. And most of the dog breeds are not pure breeds but mix breeds. This result is similar to a previous study stated that mixed breeds have higher possibilities to be adopted compared with a specific breed (Němcová & Novák, 2003). Among the popular dog breeds, the top 2 breeds are Poddle and ShihTzu. Both of them are categorized as toy breeds or lapdogs. (Protopopova et al., 2014) and (Svoboda & Hoffman, 2015) reveal that Shih Tzu is one of the most popular breeds among adopters.

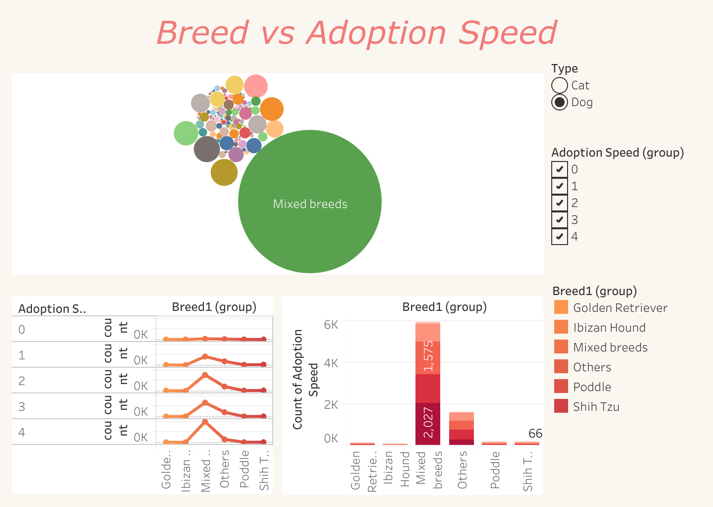
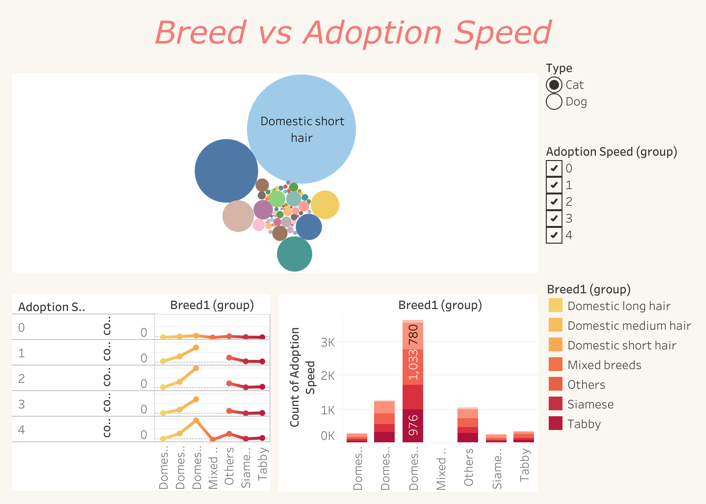


Figure 9: Breed and Adoption Speed

* ‘Color’ ‘Fee’ ‘Health’ ‘Vaccinated’ ‘Sterilized’ ‘Dewormed’ and ‘AdoptionSpeed’

Unlike previous research, which has revealed that yellow dogs have shorter stay in Shelter compared with black dogs (Kay, Coe, Young & Pearl, 2018). The most common colors are black and brown and there are almost no gray or yellow dogs. It is surprising that people prefer to adopt pets neither vaccinated nor sterilized. However, a healthy, dewormed pet tend to be adopted faster. One possible reason might be the pet age. Since most of the quickly adopted pets are less than 6 months, it is not a proper time to make pets fully vaccinated. Even though the dataset shows vaccinated pets are the minority but most of them do have deworming. This might because of veterinarian would recommend deworming pets for newborn pets, who are only 2 to 3 weeks. Lepper et al. (2012) found that pets suffering from injury problems have decreased their possibility to be adopted, probably because of the higher cost for care. Moreover, some adoptions of the injury pets may occur since adopters feel sympathy for these four-leg companions and more likely to take care of them.

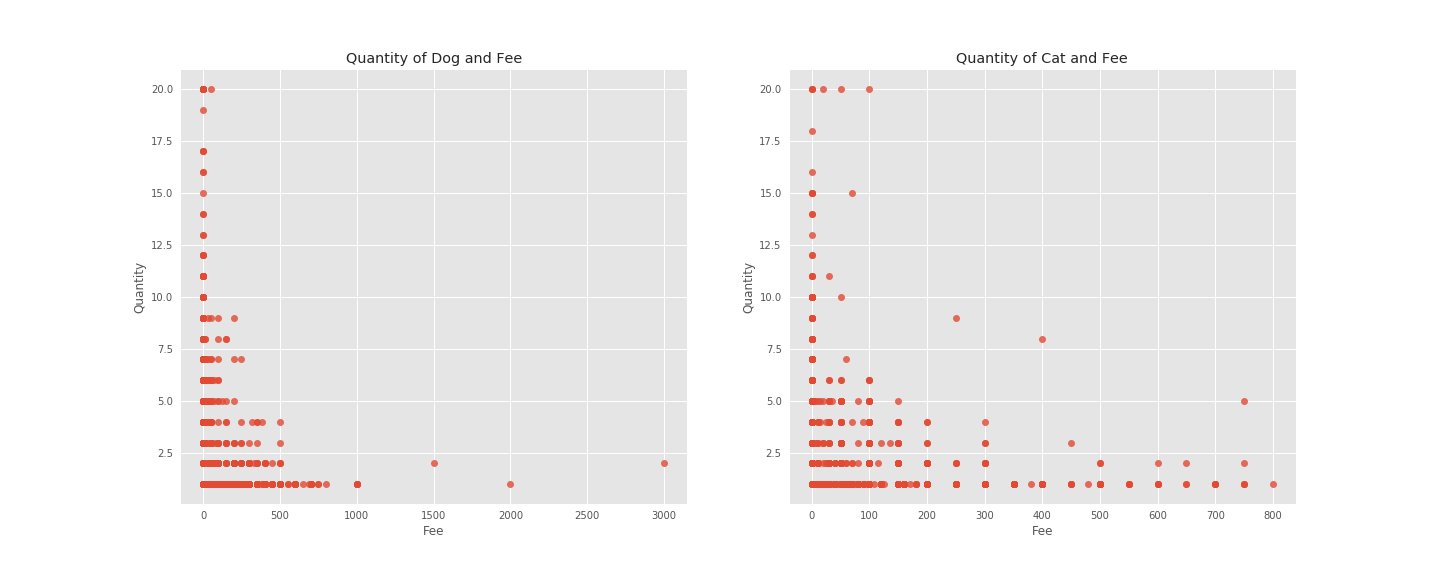


Figure 10: Dog/Cat Fee

From Figure 10 above, most of the pets are free or have a fee lower than $100 and to adopt a free dog is slower than to adopt a free cat. For some pets, as the quantity of the pets significantly lower, the fee is significantly higher. This probably because of the well-trained or some relatively expensive pure breed they are belonging to.

* ‘Name’ and ‘AdoptionSpeed’

From Figure 11 below, the popular pet names are ‘Lucky, ‘Mimi’, ‘Brownie’, etc. There are also type of pets like ‘Kitten’, ‘Puppies’ and even genders ‘Girl’, ‘Boy’.



Figure 11: Word Cloud for Name Cat/Dog

After calculation, there are around 10% of pets do not have names, however, their possibility to be adopted is higher than pets having names. One more interesting finding is that there are some names with only 1 or 2 characters, or some pets’ names are created with symbols, which have no meaning in it. There does not appear to have a direct relationship between if a pet has a name and the adoption speed.

The two models on NLP chosen for ‘Description’ variable and ‘AdoptionSpeed’ variable are Multinomial Naïve Bayes Classifier and ensemble method classification. On the first try, Multinomial Naïve Bayes classifier is applied since it provides a nice baseline for several variants of a classifier, the multinomial variant will be the one most suitable for word counts. After load the description and adoption speed into ‘X’ and ‘y’ variables respectively with train and valid, the first thing needs to do is text preprocessing, such as tokenization, lower cases, lemmatizing. Stop-words removal is included in TF-IDF vectorizer, which computes relative frequency a word appears in the description and compared the frequency across all description in the dataset. Then fit the model, and the Multinomial Naïve Bayes model resulted in a score of 35% of accuracy. Classifiers tend to have some parameters, trying all different parameter is not possible to operate, however, it is feasible to run exhaustive searches of the best parameters on a grid of values, including all classifiers with bi-grams or monogram, with or without TF-IDF, or number of parameters, or number of depth in Random Forest.

Next, to improve the accuracy, performing ensemble method machine learning on the text document and will be necessary. Similar to the Multinomial classification method, text preprocessing, such as tokenization, lower cases, lemmatizing need to be done before applying the ensemble method. Also, apply stemming and punctuation removal to the description cleaning process. For vectorizing descriptions, rather than only choosing TF-IDF, Bi-gram Count Vectorizing also added to find the best model. Ensemble methods will construct the best Random Forest model by exhaustively search across all parameter combinations. After grid searching, the best parameters should apply to random forest from Bi-gram vectorizer has the number of estimators as 300 and the maximum number of levels in each decision tree as 90. However, the best parameters from TF-IDF vectorizer as 300 of trees with no maximum number of levels in each decision tree. The following steps will be similar to the Multinomial model: split the description into ‘X’ and ‘y’, fit the model and evaluate the predictions. The accuracy score of Random Forest Classifier from Bi-gram is 42.2%. The accuracy score of Random Forest Classifier from TF-IDF vectorizer is 43%. The confusion matrix indicates that the label4, which is no pet being adopted, predicted best among other labels, following by label2, which is adopted within 1 month.

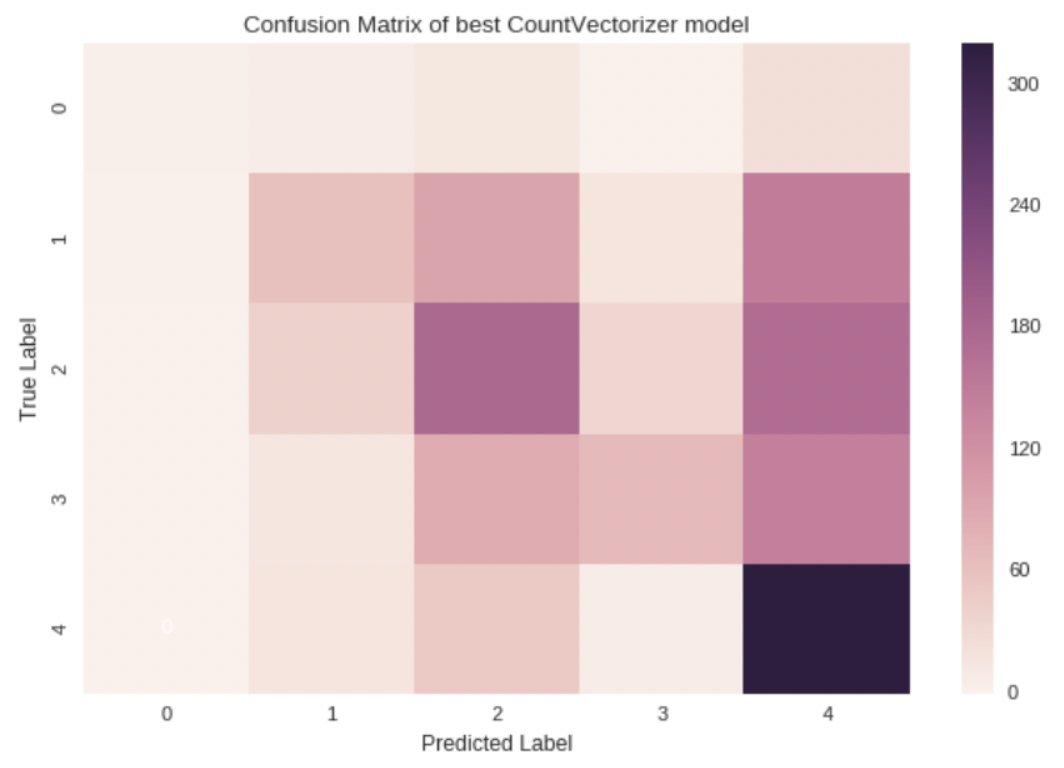
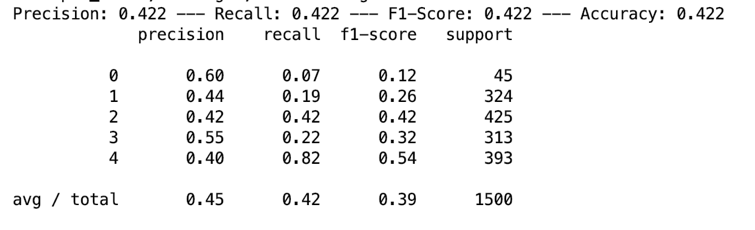


Figure 11: Bi-gram accuracy and confusion matrix

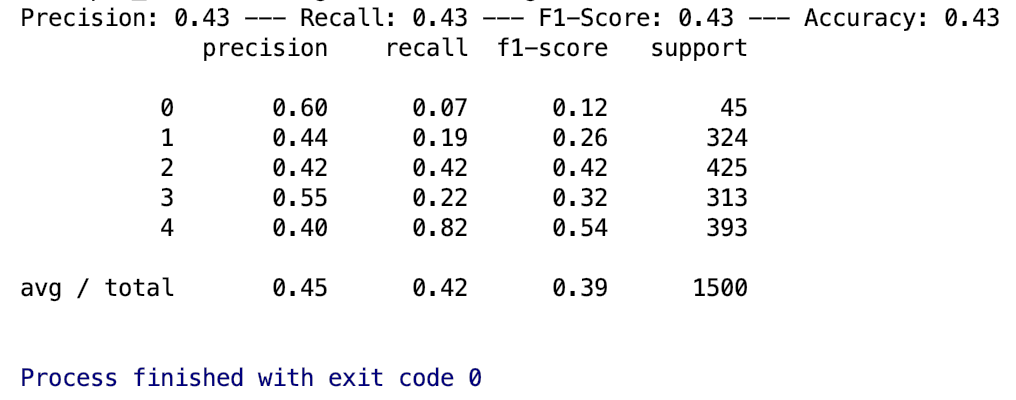
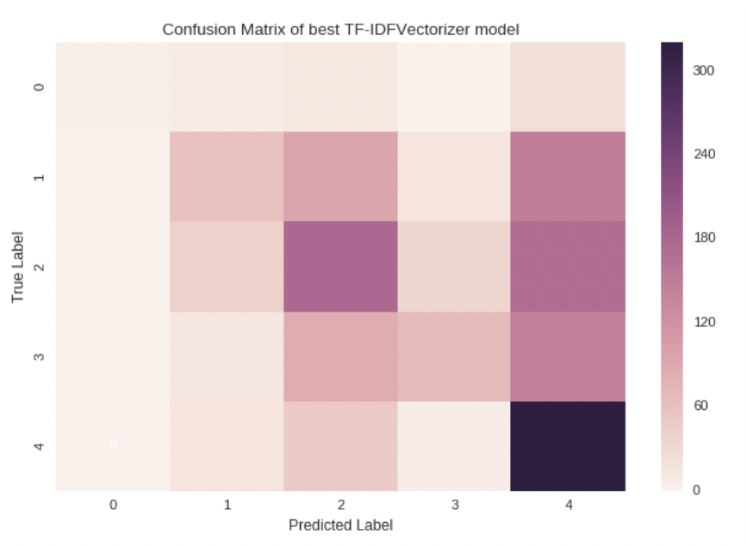
****

Figure 12: TF-IDF accuracy and confusion matrix

**RECOMMENDATIONS**

In the ‘Description’ variable, even though most of the sentences are written in English, there are still some values are written in Chinese or Malay. Separating into different language bases and do natural language processing. Another distinct method may be considered to apply to predict the adoption speed, such as using XGBoost for classifying, which follows a sequential method and compare the results to the best Random Forest model, which applies random parallel method. Since to classify 5 labels with a supervised learning method might not obtain an ideal accuracy score, another prediction could be further supported by combining the ‘AdoptionSpeed’ variable into ‘Being adopted’ and ‘Not being adopted’, and then try the same process might increase the accuracy.

**CONCLUSION**

Pet adoption and reduce euthanasia are two of the most crucial tasks for not only Word Animal Associations. With enjoyable moments owners can be obtained from owning pets, when it is necessary to relocate or having financial problems, do not relinquish the pet would be the best choice. However, as there are so many pets being abandoned, analyzing the features, which can help them find new homes are the first task. This capstone project developing data mining, visualization and machine learning on NLP attempt to find some characters of pets can increase their possibilities to be adopted. The accuracy results from NLP are low, and clearly, more data needs to be implemented in order to get better accuracy. However, from the visualization analysis, a few outcomes appear. The physical characteristics of age scaled in months, cat or dog, colors of pet, gender, maturity size, fur length, any injury, a specific breed of cats or dog, have impacts on adoption speed. Continued trend analysis with implementing another classification and improving natural language processing has the potential to provide some other support to increasing pet adoption possibilities.

**BIOGRAPHY**

**Jue Wang** is a graduate student in the Data Science Program at the George Washington University. Her interests include pet analysis and Economics behind pet adoption, financial analysis, and mathematics. She had an internship at the financial agent in Texas in 2018. She likes yoga and watching movies.

**REFERENCES**

American Society for the Prevention of Cruelty to Animals. (n.d.). Pet statistics. Retrieved from <http://www.aspca.org/about-us/faq/pet-statistics>

WeissE., MillerK., Mohan-GibbonsH., & VelaC. (2012). Why did you choose this pet? Adopters and pet selection preferences in five animal shelters in the United States. Animals, 2, 144–149. DOI:10.3390/ani2020144

Winograd, N. 2009. *Redemption: The myth of pet overpopulation and the no kill revolution in America,* 2nd ed. Los Angeles, CA: Almaden. DOI: <https://doi.org/10.1080/10888705.2013.740967>

Pet Food Manufacturers Association. (2012). *Pet population 2008 to 2012*. Retrieved from <http://www.pfma.org.uk/pet-population-2008-2012>. DOI: 10.1080/10888705.2014.883924

American Veterinary Medical Association. (2002). U.S. pet ownership & demographic sourcebook. Schaumburg, IL: Author. DOI: 10.1080/10888705.2014.971156

Coe, J. B., Young, I., Lambert, K., Dysart, L., Nogueira Borden, L., & Rajić, A. (2014). A scoping review of published research on the relinquishment of companion animals. Journal of Applied Animal Welfare Science, 17, 253–273. DOI: <https://doi.org/10.1080/10888705.2014.899910>

Salman, M., New, J. G. Jr., Scarlett, J. M., Kass, P. H., Ruch-Gallie, R., & Hetts, S. (1998). Human and animal factors related to relinquishment of dogs and cats in 12 selected animal shelters in the United States. Journal of Applied Animal Welfare Science, 1, 207–226. DOI: <https://doi.org/10.1207/s15327604jaws0103_2>

Shore, E. R. (2005). Returning a recently adopted companion animal: Adopters' reasons for and reactions to the failed adoption experience. Journal of Applied Animal Welfare Science, 8, 187–198. DOI: [10.1207/s15327604jaws0803\_3](https://doi.org/10.1207/s15327604jaws0803_3)

Posage, J. M., Bartlett, P. C., & Thomas, D. K. (1998). Determining factors for successful adoption of dogs from an animal shelter. *Journal of the American Veterinary Medical Association*, 212, 478–482. **DOI:**10.1080/10888705.2014.883924

Hart, L., Takayanagi, T. and Yamaguchi, H. 1998. Dogs and cats in animal shelters in Japan. *Anthrozoös*, 11: 157–163. Doi: <https://doi.org/10.2752/089279398787000706>

Diesel, G., Smith, H., & Pfeiffer, D. U. (2007). Factors affecting time to adoption of dogs re-homed by a charity in the UK. *Animal Welfare*, 16, 353–360. **DOI:**10.1080/10888705.2014.883924

Siettou, C., Fraser, I. M., & Fraser, R. W. (2014). Investigating some of the factors that influence ‘consumer’ choice when adopting a shelter dog in the United Kingdom. Journal of Applied Animal Welfare Science, 17, 136–147. doi: <https://doi.org/10.1080/10888705.2014.883924>

Brown, W. P., & Morgan, K. T. (2015). Age, Breed Designation, Coat Color, and Coat Pattern Influenced the Length of Stay of Cats at a No-Kill Shelter. Journal of Applied Animal Welfare Science, 18(2), 169-180. doi: 10.1080/10888705.2014.971156

PetFinder.my. (2019). PetFinder.my Adoption Prediction. Retrieved from <https://www.kaggle.com/c/petfinder-adoption-prediction>

ZONODO

O’Connor, R., Coe, J. B., Niel, L., & Jones-Bitton, A. (2016). Effect of Adopters’ Lifestyles and Animal-Care Knowledge on Their Expectations Prior to Companion-Animal Guardianship. Journal of Applied Animal Welfare Science, 19(2), 157-170. doi: [10.1080/10888705.2015.1125295](https://doi.org/10.1080/10888705.2015.1125295)

Brown, W. P., Davidson, J. P., & Zuefle, M. E. (2013). Effects of phenotypic characteristics on the length of stay of dogs at two no kill animal shelters. *Journal of Applied Animal Welfare Science*, 16, 2–18. doi:10.1080/10888705.2013.740967

DeLeeuw, J. L. (2010). *Animal shelter dogs: Factors predicting adoption versus euthanasia* (Unpublished doctoral dissertation). Wichita State University, Wichita, Kansas, US. doi: 10.1080/10888705.2018.1436917

Diesel, G., Smith, H., & Pfeiffer, D. U. (2007). Factors affecting time to adoption of dogs re-homed by a charity in the UK. *Animal Welfare*, 16, 353–360. doi: <https://doi.org/10.1080/10888705.2018.1436917>

Lepper, M., Kass, P. H., & Hart, L. A. (2002). Prediction of adoption versus euthanasia among dogs and cats in a California animal shelter. *Journal of Applied Animal Welfare Science*, 5, 29–42. doi:10.1207/S15327604JAWS0501\_3

Žák, J., Voslářová, E., Večerek, V., & Bedáňová, I. (2015). Sex, age and size as factors affecting the length of stay of dogs in Czech shelters. *Acta Veterinaria Brno*, 84, 407–413. doi:10.2754/avb201584040407

Protopopova, A., Mehrkam, L. R., Boggess, M. M., & Wynne, C. D. L. (2014). In-kennel behavior predicts length of stay in shelter dogs. *PLoS One*, 9, e114319. doi: 10.1371/journal.pone.0114319

Svoboda, H. J., & Hoffman, C. L. (2015). Investigating the role of coat colour, age, sex, and breed on outcomes for dogs at two animal shelters in the United States. *Animal Welfare*, 24, 497–506. doi:10.7120/09627286.24.4.497

Němcová, D., & Novák, P. (2003). Adoption of dogs in the Czech Republic. *Acta Veterinaria Brno*, 72, 421–427. doi:10.2754/avb200372030421

Kay, A., Coe, J. B., Young, I., & Pearl, D. (2018). Factors Influencing Time to Adoption for Dogs in a Provincial Shelter System in Canada. Journal of Applied Animal Welfare Science, 1-14. doi: [10.1080/10888705.2018.1436917](https://doi.org/10.1080/10888705.2018.1436917)

**APPENDIX A**

|  |  |  |
| --- | --- | --- |
| Feature Variables | Data Type | Description |
| PetID | object | Unique hash ID of pet profile |
| AdoptionSpeed | int | Categorical speed of adoption. Lower is faster. |
| Type | object | Type of animal (1: dog, 2: cat) |
| Name | object | Name of pet |
| Age | int | Age of pet when listed, in months |
| Breed1, Breed2 | object | Breed1: primary breed of pet,  Breed2: secondary breed of pet, if pet is mixed breed |
| Gender | object | Gender of pet (1: male, 2: female, 3: neutered/spayed) |
| Color1, Color2, Color3 | object | Color1: the first color of pet,  Color2: the second color of pet,  Color3: the third color of pet |
| MaturitySize | object | Size at maturity (1: small, 2: medium, 3: large, 4: extra-large, 0: not specified) |
| FurLength | object | Fur Length (1: short, 2: medium, 3: long, 0: not specified) |
| Vaccinated | object | Pet has been vaccinated (1: yes, 2: no, 3: not sure) |
| Dewormed | object | Pet has been dewormed (1: yes, 2: no, 3: not sure) |
| Sterilized | object | Pet has been spayed/neutered (1: yes, 2: no, 3: not sure) |
| Health | object | Health condition (1: health, 2: minor injury, 3: serious injury, 0: not specified) |
| Quantity | int | Number of pets represented in profile |
| Fee | int | Adoption fee (0: free) |
| Description | object | Profile write-up for this pet. The primary language used is English, with some in Malay or Chinese. |
| DataType | object | The data is from train or test |

**APPENDIX B: variable ‘AdoptionSpeed’ description**

|  |  |
| --- | --- |
| Value | Description |
| 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. |