Data Mining for Business (BUDT758T)

Project Title: Explaining features affect pet adoption decision

Team Members: Tianjie Xu

Guiran Niu

Zhongchuan Xiao

Jiajie Tang

Tianxing Liang

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
Contact Author	Tianxing Liang	
	Zhongchuan Xiao	
	Guiran Niu	
	Tianjie Xu	
	Jiajie Tang	

I. Executive Summary

"Friends Furever", the most shared branded video of 2015, is literally nothing but animals being cute. It has been watched nearly 22 million times on YouTube and shared 6.5 million times on social media. A study of Youtube videos shows that either 30 or 40 percent of the most-shared videos had the animal theme in common. However, millions of stray animals suffer on the street or are euthanized in shelters. On the website of PetFinder (data provider), there are more than 8000 homeless dogs and 6500 neglected kitties. If a bridge between pet lovers and shelters can be built faster, many lives could be saved and more happy families will be created. How adorable is a small miracle in the shelter so a person is willing to adopt it? In other words, what features will impact the adoption speed of a pet? We can give you some answers.

Age is one of the most important influential factors. People prefer to adopt more matured animals because older means a higher survival probability in general. The length of the description is as important as the number of photos in the pet's profile, what's more, the more negative words in a description, the faster a pet will be adopted because negative words can trigger the sympathy inside people. One last interesting finding is a sterilized pet has a lower odd of adoption maybe because median of age is only 3 months which is not good for a pet to be sterilized. By using principal findings, the shelter will know how to finalize pets' profile to increase adopt rate or develop a proper solution for pets with some certain features.

II. Data Description

Data Source:

PetFinder.my Adoption Prediction from Kaggle (https://www.kaggle.com/c/petfinder-adoptionprediction)

Department of Statistic Malaysia (www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=455&bul_id=SnJlWjNGZ3VWajUraDlBcFpMQ3JWUT09&menu_id=U3VPMldoYUxzVzFaYmNkWXZteGduZz09)

Pet breed (https://dogtime.com/dog-breeds; https://cattime.com/cat-breeds)

Data preprocess:

State data: In the origin data set, the 'state' is represented by its ZIP code in Malaysia, which is meaningless in the model. To better explain the geographic factors, we add three variables may be helpful in our project: Median income, average house size, and crime index.

Pet features: Based on the variable 'Breed ID' in original data, we picked 11 characteristics that best describe cats and dogs. For each unique breed ID, we scraped the corresponding ratings of the 11 characteristics from the same website.

PCA1-5: After adding 11 pet features from other websites, these variables may be highly correlated when used in our models, such as 'general health', 'exercise needs' and 'life expectancy'. We used PCA to convert these possibly related

variables into a set of values, 'PCA1' to 'PCA5', as the first five components account for or explain more than 92% of the overall variability.

Sentiment: By doing sentiment analysis on the description in R ('sentimentr' package), we get the sentiment score and regard them as an important variable to check whether people will be influenced by the emotion contains in others' description. Also, the number of words in each description is also taken into account to reflect the detail abundance of description.

Sample Size： Total 14990 observations, in which 10494 (70%) of them are segmented into training data and 4496 (30%) of them are treated as testing data.

Number of Variables: Finally, 44 variables. In which 18 of them are categorical variables, and 26 of them are numerical variables.

Our original dataset from Kaggle contains only the profile elements of pets. However, there are tons of factors exist but not included in this data set that can change people's mind on adoption. We started from brainstorming characteristics of different breeds because based on our common sense people would like to adopt pets that are easy to groom, smart and friendly. We also considered state demographic factors that could accelerate adoption speed. For example, people may want to adopt a dog if they live in an area with a higher crime index or richer people may influence adoption speed since they are more affordable to adopt a new pet. In the end, we also want to see whether the sentiment of description in pet profile and the length of description can affect the adoption speed because sometimes people may want to know more about the pet from description before they decide to adopt. We are interested in not only prove the common sense but also more on the inverted result, which is more interesting than the proof of common sense.

Data Sample:

Type	Name	Age	Breed1	Breed2	Gender	Color1	Color2	Color3	MaturitySi	FurLength	Vaccinatec	Dewormec	Sterilized	Health	Quantity
2	Nibble	3	299	0	1	1	7	0	1	1	2	2	2	1	1
2	No Name '	1	265	0	1	1	2	0	2	2	3	3	3	1	1
1	Brisco	1	307	0	1	2	7	0	2	2	1	1	2	1	1
1	Miko	4	307	0	2	1	2	0	2	1	1	1	2	1	1
1	Hunter	1	307	0	1	1	0	0	2	1	2	2	2	1	1

Fee	State	RescuerID	VideoAmt	Descriptio	PetID	PhotoAmt	AdoptionS	Dog Frienc	Easy To Gr	Easy To Tr	Exercise N	Friendly T	General H	Incredibly	Intelligenc
100	41326	8480853f5	0	Nibble is a	86e1089a3	1	2	3	0	0	3	0	0	3	5
0	41401	3082c7125	0	I just foun	6296e909a	2	0	3	0	0	3	0	0	3	5
0	41326	fa90fa5b1	0	Their preg	3422e4906	7	3	3.531707	3.356098	3.604878	4.141463	3.458537	3.390244	4.136585	4.082927
150	41401	9238e4f44	0	Good guar	5842f1ff5	8	2	3.531707	3.356098	3.604878	4.141463	3.458537	3.390244	4.136585	4.082927
0	41326	95481e953	0	This hands	850a43f90	3	2	3.531707	3.356098	3.604878	4.141463	3.458537	3.390244	4.136585	4.082927

Potential F	Tendency	life_span	mixbreed	Crime Ind	living spac	Median in	adoptions	most popu	PC1	PC2	PC3	PC4	PC5	word_cou	ave_sentir
3	3	16	0	408.6	1774	7225	0	0	-3.02721	0.348039	-0.07053	-0.0492	0.105639	68	0.430272
3	3	16	0	716.9	827	9073	1	0	-3.02721	0.348039	-0.07053	-0.0492	0.105639	23	0.262513
4.385366	3.20098	13.80976	1	408.6	1774	7225	0	0	2.145564	-0.24741	0.058694	-0.04811	-0.06547	69	0.15016
4.385366	3.20098	13.80976	1	716.9	827	9073	0	0	2.145564	-0.24741	0.058694	-0.04811	-0.06547	25	0.84
4.385366	3.20098	13.80976	1	408.6	1774	7225	0	0	2.145564	-0.24741	0.058694	-0.04811	-0.06547	79	0.439272

Data Description:

Variable	Description	Type
Dependent Variable		
AdoptionSpeed	*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. (There are no pets in this dataset that waited between 90 and 100 days).*	*Categorical-Integer*
Adopted	*0-No adoption after 100 days of being listed.*	*Categorical-Integer*
Original Data		
Type	1-Pet is dog 2-Pet is cat	Categorical-Integer
MixedBreed	0-Pure1-Mixed bread	Categorical-Integer
Breed1	Primary breed of pet	Categorical-Integer
Breed2	Secondary breed of pet, if pet is of mixed breed	Categorical-Integer
Gender	Gender of pet (1 = Male, 2 = Female, 3 = Mixed, if profile represents group of pets)	Categorical-Integer
Color1	Color 1 of pet	Categorical-Integer
Color2	Color 2 of pet	Categorical-Integer
Color3	Color 3 of pet	Categorical-Integer
Age	Age in months	Continuous-Integer
MaturitySize	Size at maturity (1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large, 0 = Not Specified)	Categorical-Integer
FurLength	Fur length (1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified)	Categorical-Integer
Vaccinated	Pet has been vaccinated (1 = Yes, 2 = No, 3 = Not Sure)	Categorical-Integer
Dewormed	Pet has been dewormed (1 = Yes, 2 = No, 3 = Not Sure)	Categorical-Integer
Sterilized	Pet has been spayed / neutered (1 = Yes, 2 = No, 3 = Not Sure)	Categorical-Integer
Health	Health Condition (1 = Healthy, 2 = Minor Injury, 3 = Serious Injury, 0 = Not Specified)	Categorical-Integer
Quantity	Number of pets represented in profile	Continuous-Integer
Fee	Adoption fee (0 = Free)	Continuous-Integer
State	State location in Malaysia	Categorical-Integer
RescuerID	Unique hash ID of rescuer	Categorical-Integer
VideoAmt	Total uploaded videos for this pet	Continuous-Integer
PhotoAmt	Total uploaded photos for this pet	Continuous-Integer
Description	Profile write-up for this pet. The primary language used is English, with some in Malay or Chinese.	Categorical-Integer
State Data		
Median income	Median income by state	Continuous-Integer
Average house size	The average sizes of new residential launches among the Malaysian states in 2017	Continuous-Integer
Crime index ratio per 100,000	Crime index ratio per 100,000 population for Malaysia by state in 2017	Continuous-Integer
Pet Feature		
Dog Friendly	Friendship to other dogs, ranked from 0 to 5 while 0 represent hostile and 5 represent friendly.	Continuous-Integer
Easy To Groom	Whether behave well when being groomed, ranked from 0 to 5 while 0 represent choleric and 5 represent well behaved.	Continuous-Integer
Easy To Train	Whether behave well when being trainned, ranked from 0 to 5 while 0 represent incorrect response and 5 represent correct response.	Continuous-Integer
Exercise Needs	Need of exercise, ranked from 0 to 5 while 0 represent low needs and 5 represent high needs.	Continuous-Integer
Friendly Toward Strangers	Friendship to strangers, ranked from 0 to 5 while 0 represent hostile and 5 represent friendly.	Continuous-Integer
General Health	This breed general health situation, ranked from 0 to 5 while 0 represent sick and 5 represent healthy.	Continuous-Integer
Incredibly Kid Friendly Dogs	Friendship to kids, ranked from 0 to 5 while 0 represent aggressive and 5 represent friendly.	Continuous-Integer
Intelligence	Average intelligence of this breed, ranked from 0 to 5 while 0 represent stupid and 5 represent intelligent.	Continuous-Integer
Potential For Playfulness	Tendency to play with you , ranked from 0 to 5 while 0 represent independent and 5 represent dependent.	Continuous-Integer
Tendency To Bark Or Howl	Tendency to bark, ranked from 0 to 5 while 0 represent silent and 5 represent noisy.	Continuous-Integer
LifeSpan	Maxium life span of this breed.	Continuous-Integer
PC1	The first PCA conponent of above pet features, proportion of variance=0.600, cumulative proportion=0.600	Continuous-Float
PC2	The second PCA conponent of above pet features, proportion of variance=0.1164, cumulative proportion=0.7164	Continuous-Float
PC3	The third PCA conponent of above pet features, proportion of variance=0.09037, cumulative proportion=0.80679	Continuous-Float
PC4	The fourth PCA conponent of above pet features, proportion of variance=0.06381, cumulative proportion=0.87060	Continuous-Float
PC5	The fifth PCA conponent of above pet features, proportion of variance=0.05539, cumulative proportion=0.92598	Continuous-Float
Text Processing		
word_count	The number of words in description.	Continuous-Integer
ave_sentiment	The sentimet score of description.	Continuous-Float
Data Processing		
most popular	Whether this breed is popular among adoption. 1 if half of this breed is adopted in 100 days. 0 otherwise.	Data Processing

III. Research Questions

The number of dogs and cats euthanized in U.S. shelters annually has declined from approximately 2.6 million in 2011. This decline can be partially explained by an increase in the percentage of animals adopted and an increase in the number of stray animals successfully returned to their owners. This situation may be worse in Malaysia. On the website of PetFinder, this project will help those shelters in Malaysia better allocate their resources to help pets to be adopted and reduce the euthanasia.

As an animal shelter agency, the adoption rates and adoption speed are the key concerns of the PetFinder. By increasing the adoption rates and reducing the adoption waiting time, the PetFinder can not only alleviate the homeless pets' suffering for waiting the adoption but also reduce its carrying cost of these unadopted animals.

The adoption behavior of the potential adopters could be affected by many factors. People make the adoption decision based on both rational and emotional reasons. In this project, we will use the pattern that bearded in the historical data to investigate what kinds of factor will trigger the adoption decision of potential adopter, which will result in the increasing of adoption rates.

Given the data we have, in general, we will conduct research on three aspects: First, how do the features of the pet itself affect the adoption rate? These features will include both physical factors such as the age and the perceptual factors such as the "friendliness level" of the animal. Second, how does the way that the PetFinder list the pet affect the adoption rate? These features are controllable by the PetFinder, such as the length and sentiment of the description, the amount of Photo of the pet, and the number of animals listed in one profile. Third, how do the demographic factors of shelters affect the adoption rate? These factors will be crime index, the per capita living space, and median income by State.

IV. Methodology

KNN: The K-Nearest Neighbor Algorithm (*KNN*) is based on a simple underlying idea: to classify a case based on those similar cases that in the train data. And intuitively, we think this base idea is sure to be usefully in our prediction for most people tend to have similar preferences and therefore tend to show their love to some similar pets. And KNN is pretty good at handling a local issue, which is very likely to happen in our study.

Neural Network: Neural network is a very effective learning algorithm. Although we cannot get a meaningful interpretation of estimated parameters, by capturing the complicated relationship and giving a good predictive performance, Neural network is still widely used for classification and prediction. In our data set, most variables are categorical, and our goal is a classification result: Adoption Speed. So Neural Network is a good model for our data set.

Naive Bayes is a good classification method based on the probability model which can handle both categorical and numerical variables. Most of the variables in the original data set are categorical variables which suit this model perfectly. We can also find out the attributes that are valuable and significant to the dependent variables. It also provides some insights into those variables by comparing the probabilities of each value and calculate the probabilities for each class.

Classification tree and Ensemble methods(Bagging, Random Forest, Boosting): Based on our goal of discovering the factors that affect the speed of pet adoption, classification trees will be the most intuitive model for classification and profiling. This methodology will give us a nice visualization to interpret our classified profile without variable selection.

Since we have categorical and continuous variables and considering the fact that we may have outliers, this methodology is also a good fit for this situation. Meanwhile, decision tree will constantly be the one to help us explore the data during the process. To best trim the model and get better accuracy, we introduced ensemble methods with the classification tree to learn the patterns in data using different models and combine this learning.

LGM: The logistic regression model will not only provide us the predicted classification but also show us how factors affect the probability of the outcome. In our case, one of our research questions is to investigate how the factors we collected in the data influence the adoptions of animals. To answer this question, we applied the logistic regression model. In the model, the "Adopted" has been used as the dependent variable. Other 29 variables include both categorical and numerical were used as predictors.

PCA: After adding pet features we collected from website, the total number of variables comes to more than 40. Principle sentiment analysis is a useful dimension-reduction tool that can help to transform a number of (possibly) correlated variables into a smaller number of uncorrelated variables. This may lead to a loss of some information but improve our models.

Sentiment Analysis: The 'Description' in the original data set is one of the valuable variables, as a long and positive description can help the pets be adopted quickly. By adding 'wordcount', the number of words in the description is not the best way to make good use of this variable, sentiment analysis can quantify how positive these words are.

V. Results and Findings

We tuned the models by 7 steps following the thought of "Forward Stepwise Feature Selection".

In the first step, basic features of pets were used such as age, gender, dewormed or not, how much is adoption fee, and how many photos were attached for a pet. Based on the above features, mix breed and how long is the description were added. However, accuracies were not decent for all of models using adoption speed 0 to 4 as the dependent variable. To improve accuracy, features selected above were constant, the dependent variable classified as 0 and 1 in step 2. 0 means the pet is not adopted, 1 means the pet is adopted. In step 3, based on basic and added features from the first step, the crime index, the per capita living space, and median income by State were added to present demographic factors. Because there were so many breeds in data set, a factor called "most popular" was created by whether a breed was adopted in 100 days or not in step 4. The average sentiment was added step 5 to put sentiment of description into models. In step 6, pet features such as "friendly", "Easy to Groom" were added. PCA was used to reduce the factors from "pet feature" in the last step.

Logistic Regression Model:

After different combination of aggregated variables mentioned above, for the logistic regression model, we find out that the original data set with the three demographic variables of location, the "most popular" binary variables, the sentiment variables and the 11 pet personality variables will give the highest prediction accuracy, which is 0.7459275. The confusion matrix of this model is shown in Exhibit 1

Because of the nature of the data, out of 4113 testing data, 73% of them are with label “adopt=1”, and 27% of them are with label “adopt=0”. In this case, the sensitivity of this model is 0.19823 and the specificity is 0.9534. This logistic model will most likely to correctly predict a pet that will be adopted within 100 days, but on the other hand, it will not give a very accurate prediction for the pet that will not be adopted within 100 days.

The logistic regression model will also exhibit the significant level of each variable has on the dependent variables. According to the result, Age, Color, FurLength, Vaccinated, Dewormed, Sterilized, Quantity, Fee, Photo amount, Easy to train, Exercise Needs, General Health, mix breed, most popular and Median income by state are the variables that have high significant level. In other words, these factors have most effect on the adoption rate. The detailed result is shown in Exhibit 2.

With this result, we can answer the three main research question for this project. For the first one, how the features of the pet itself affect the adoption rate and speed? Based on the logistic regression model, Age is one of the most important one. 1-month increase in age will increase the odds of adoption by a factor of 2.341988. This result indicates that people will tend to adopt the pet that is relatively older. Considering that the age of pet is counted in months, we can make the assumptions that people tend to avoid adopting the animal that is in the suckling period, which makes them more vulnerable and needs more care. This finding shows that for the shelters, they need to do more preparation for caring the animal in the infancy period since in most case they will not be adopted until they are older.

The result indicates that the hygiene of the pet is also an important consideration for the adoption. If a pet is labeled as “Dewormed=2” which indicate that it is not been dewormed, the odds of it be adopted will decrease by a factor of 0.762007. It is accord with the intuition that for the animal in the shelter, people will be sensitive to its hygiene status. An animal has not been dewormed will raise people’s doubt for its’ health and hygiene. In order to ease people’s doubt and increase the rate of adoption, the shelters should do the deworm for all animals in the shelters.

Among the 11 characteristic factors of pets, two of them exhibit a strong effect on the adoption rate, which is “Easy to train” and “Exercise needs”. Both of them are in line with intuition: People will tend to adopt pet that is easy to train, and the pet that does not require too much exercise.

For our second main research question: how does the way that the PetFinder list the pet affect the adoption rate and speed? The logistic regression model indicates three factors will have a high influence on the adoption rate. The first one is the “Quantity”, which is the number of pets listed in one profile. In most case, only one animal will be listed in one profile, but in some case a litter of pets will be listed together in one profile, and the adopter will have to adopt them all at once. The result indicates that 1 unit increase in the quantity of pets decreases the odds of adoption by a factor of 0.85274. It is very intuitive to think that to adopt more pets, people will need more resources and higher conditions, for most people they are not capable of adopting multiple animals at one time. The second factor is the Fee. It is also intuitive to conclude that for an animal with less adoption fee, the odds of its being adopted will increase. The third factor that is important to the adoption rate, and can be considered an interesting one, is the amount of photo that

uploaded for the pet. The logistic regression model proves that the visual demonstration of the pet will help to increase the adoption rate. Based on the result, we can conclude that by having more photos about the pet, people will have a more perceptual cognition of it, and trigger their desires to adopt it. These three factors can be considered as the most important findings we get from the logistic regression result since all of them can be directly controlled by the PetFinder. Our findings clearly provide three recommendation for the PetFinder to do to increase the adoption rate: First, the PetFinder should decrease the number of pets that listed in one profile. It should try to split up the profile with multiple pets. Second, the PetFinder should decrease the adoption fee if it wishes to increase the adoption rate. Third, it should provide more photo for the animal listed in the profile.

For our last research question: "how the features of the location of shelters affect the adoption rate and speed?" The result shows that the only factor that has a significant influence on the adoption rate is the "Median.income.by.State", which exhibit a very intuitive pattern states that the shelters that located in the wealthy state will have higher adoption rate.

There is one counter-intuitive result for the features of the animal. According to logistic regression model, people will more willing to adopt the animal that is not sterilized. The odds of the animal be adopted will increase by a factor of 2.282109 if the animal is not be sterilized. Based on our assumption, the sterilization, just like the deworming, can be considered a frequently applied process to make the pet healthier and more suitable for keeping as a pet. However, the sterilization is also a controversial and special procedure since it deprives pet's ability to procreate. Our finding shows that perhaps people do not want their pet to lose the ability to procreate. For the shelters, this finding clearly indicates that they should not do the sterilization if it is unnecessary or offer the sterilization as an option rather than a precondition.

Classification tree and ensemble methods:

The first step we calculated each variable's p-values to evaluate the significance of each variables in differentiating between the five levels of adoption speed, but the result shows all the variables are statistically insignificant in differentiating the levels and tend to differentiate between two levels only. The pruned tree with 5 levels are shown in Exhibit 3. The best model in the step 1 is random forest with an accuracy of 40% (Exhibit 4). Therefore, our main research of dependent variable is classified with our standard two classes, starting from step 2.

According to the 36 models of classification trees and ensemble methods, all models improved accordingly with the data gathering progress, except the pruned trees, which has only 'Age' as the split and consider everything to be the class '1' (Exhibit 5). The accuracy of the pruned tree stays at 72.13% every time because the actual class '1' contains the majority of the observations and the model couldn't find the pattern of the dependent variable and its predictors due to the limitation of the variables itself. The tree simply ignores the predictors shown in Exhibit 6.

Overall, the best model with the highest accuracy of 76.47% is random forest with either pet features or the PCA of pet features in step 6 or 7 (Exhibit 7). This tie looks reasonable and shows our 5 PCA variables are accurate enough to be the

alternative of 11 pet features. Similarly, the best bagging model 75.87% and the best boosting model 74.29% are also with either pet features or the PCA of pet features in step 6 or 7. Since there are certain number of predictors are correlated in our dataset, random forest out-performed bagging in all the steps by decorrelation. However, it is also interesting to see bagging did well on every step than boosting. In the end, ensemble methods all did well than the tree by decreasing variance and bias. Based on our ideas of gathering outside data, we improved our best random forest model from the original 74.31% to 76.47% and most of the important variables in the best model are coming from the variables we add in (Exhibit 8). The implementation of ensemble methods also improved our model from 72.13% to 76.47%.

To better interpret our result, the variables importance plot of the two Random forest models have same top 10 variables, which are age, sterilized, character length of the description, word count of description, amount of photo in the profile, dewormed, maturity size, crime index, state income, living space, fee and sentiment of description (Exhibit 9 MeanDecreaseAccuracy). We can see that people are considering these factors the most in adoption, which are all reasonable and solved the main question of our research. Other related questions can also be solved by this model. The 3 demographic factors, which are crime index, income and living space, are all important. Among the 36 models, age is the always top 1 criteria affect adoption speed. It also proved that the adoption can be controlled by the pet profile because the sentiment and the description length of pet profile are significantly impact the adoption.

Surprisingly, the number of photos in the profile ranked in the top 5 important predictors, but the number of videos in the profile are not important at all. Another interesting finding is the characteristics of pet breed and the health of pet aren't the crucial factors compare to the age, description and photo amount in the profile. Therefore, the adoption center could worry less about some pet breeds that are not rated well on certain characteristics and whether an injured pet will affect adoption. The story from the description of the pet profile are weighted much more than physical defect of a pet.

KNN:

At first, our dependent variable is split as 5 factors, represent different speed of adoption. And the accuracy rate of KNN model in this circumstance is 0.3207. The result seems not so beautiful, which indicates that it's not so easy to accurately predict the speed.

The KNN model performs best on our issue. When we combine our dependent variables into two groups to predict whether a pet will be adopted. The KNN model reaches its best accuracy rate at 0.8323. When more variables are added into in following steps, to our surprise, the accuracy dropped a little. But compared

to use the 11 pet features as predictors, when PCA included to replace raw features' data, the accuracy rate increased a little. This is an interesting phenomenon, which indicates that not in all situations, more predictors result in better prediction result. So, try to simplify the model is a very important issue in model building, not only for computational efficiency but also for higher accuracy.

And in the process, the best k is also always changing, indicates the importance to split a validation set to find out the best k.

Neural Network:

The goal of using Neural network model is to get a decent accuracy rate.As mentioned at the beginning, the model was run by seven steps. In the first step, the accuracy is only 0.2326 which was not good. By following the instruction of step 2 to step 7, the accuracy rates were 0.7360, 0.7405,0.7363,0.7358,0.7409, and 0.7403.

From above accuracies, the third, sixth and seventh models provide good results around 0.74. By considering involving more features of pets, either sixth or seventh model will be a fit one for the problem.

Naive Bayes:

The accuracy table (Exhibit 12) shows that the accuracy has improved significantly by combining the dependent variables into 0 and 1 after step 2. The successive models all reach an accuracy of nearly 70% compared to 34% in model 1. But we can also see that the accuracy does not always increase when adding variables gradually. Even when adding the 11 breed features variables in step 6, the accuracy dropped to near 2%, since these 10 variables may be highly correlated, and the breed attributes cannot reflect the individuals' personalities well.

To get more interpretation of numerical variables, transferring those numeric variables to categorical variables by firstly scaling them to (0,1) and then bin them to 5 categories with 0.2 step each and get the result table 2. Some accuracies improve a little, while others fall. After scaling the variables, some valuable categories may be too "rare" and has biased probability estimates of class. (Exhibit 11)

Sterilized or not has the most different probabilities between two adoption speed. Among pets that are adopted, 30% of them are sterilized and 50% are not, while for pets not adopted, only 17% are sterilized.

VI. Conclusion

The highest accuracy model:

	Model with highest accuracy	Highest Accuracy
Step1	Random forest	0.4000
Step2	KNN	0.8323
Step3	KNN	0.8225
Step4	KNN	0.8185
Step5	KNN	0.8185
Step6	KNN	0.8083
Step7	KNN	0.8147

When the dependent variable was Adoption Speed, which contains 5 factors representing different adopted speed, the accuracy rate for all models are very low. The best model is Random Forest and its accuracy rate reaches 0.40. Not so pretty good but better than the baseline 0.27 (the class '4' appears the most frequent and takes up 0.27 in data).
When the dependent variable was combined into two groups to predict whether a pet will be adopted or not, the accuracy rate increased significantly to at least 0.7. KNN becomes the best model with 0.83 accuracy rate. The reason may be its advantage at finding the local pattern.
But because KNN model is not useful in interpretation and cannot reveal us the importance and correlation of predictors. We further turned to logistic regression and tree for a better explanation.

High accuracy models with meaning features:

In summary, the logistic regression model and the Classification tree model provide a global picture of how do the external factors affect the adoption rate of pets. By combining the results and findings of these two models, we are able to reach the following conclusions:

1. Age is one of the most important influential factors. The increase in the age in term of months will increase the odds of adoption. Since people tend to adopt matured animals, the shelters need to prepare for caring for the animal in the infancy period.
2. The number of photos in the pet's profile will help to increase the odds of adoption. The visual demonstration will trigger people's desires to adopt the pet. The shelters should upload more photos to the pet's profiles.
3. The hygiene of the pet is an important consideration for the adoption. The dewormed pet will have higher odds of adoption.
4. The higher the median income by state can increase the adoption rate. Therefore, the shelter can target higher income level people for higher adoption rate.
5. The sentiment and length of the profile description will highly impact the adoption rate. Since the more negative words in a description, the faster a pet will be adopted because negative words can trigger the sympathy inside people, the shelter should pay special attention when writing the profile description of each pet.
6. The sterilization procedure will have a counter effect on the adoption rate. The sterilized pet will have lower odds of adoption. Based on this result, the shelters should not actively conduct the sterilization procedures on pets.
7. The number of photos in the profile will impact the adoption rate, whereas the number of videos will not.
8. The pet features and health situation will not impact the adoption rate.
9. The story from the description of the pet profile is weighted much more than physical defect of a pet.

VII. Appendix

Exhibit 1

Actual.v	Predicted 0	1
0	224	906
1	139	2844

Exhibit 2

Variables	*Pr(>|z|)*	*Estimate*	*EXP*
Age	**< 2e-16**	**0.851**	**2.341988**
FurLength3	**0.000269**	**0.504**	**1.655329**
Dewormed2	**0.000351**	**-0.2718**	**0.762007**
Sterilized2	**< 2e-16**	**0.8251**	**2.282109**
Quantity	**3.13E-11**	**-0.1593**	**0.85274**
Fee	**0.000834**	**-0.001134**	**0.998867**
PhotoAmt	**4.71E-16**	**0.07175**	**1.074387**
Easy.To.Train	**2.27E-08**	**0.361**	**1.434763**
Exercise.Needs	**1.85E-04**	**-0.3725**	**0.68901**
mixbreed1	**2.20E-05**	**-0.3076**	**0.735209**
Median.income.by.State	**1.53E-04**	**0.000255**	**1.000255**
most.popular1	**3.12E-11**	**0.7418**	**2.099712**

Exhibit 3

Predict	0	1	2	3	4
Actual					
0	0	0	55	0	62
1	0	0	512	0	395
2	0	0	676	0	537
3	0	0	518	0	470
4	0	0	330	0	942

Exhibit 4

Predict	0	1	2	3	4
Actual					
0	2	41	31	11	32
1	1	298	299	81	228
2	0	236	457	190	330
3	2	149	302	238	297
4	0	124	235	109	804

Exhibit 5

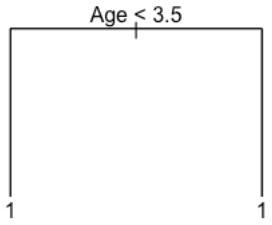

Exhibit 6

Predict	0	1
Actual		
0	0	0
1	1272	3225

Exhibit 7

Predict	0	1
Actual		
0	430	842
1	216	3009

Exhibit 8

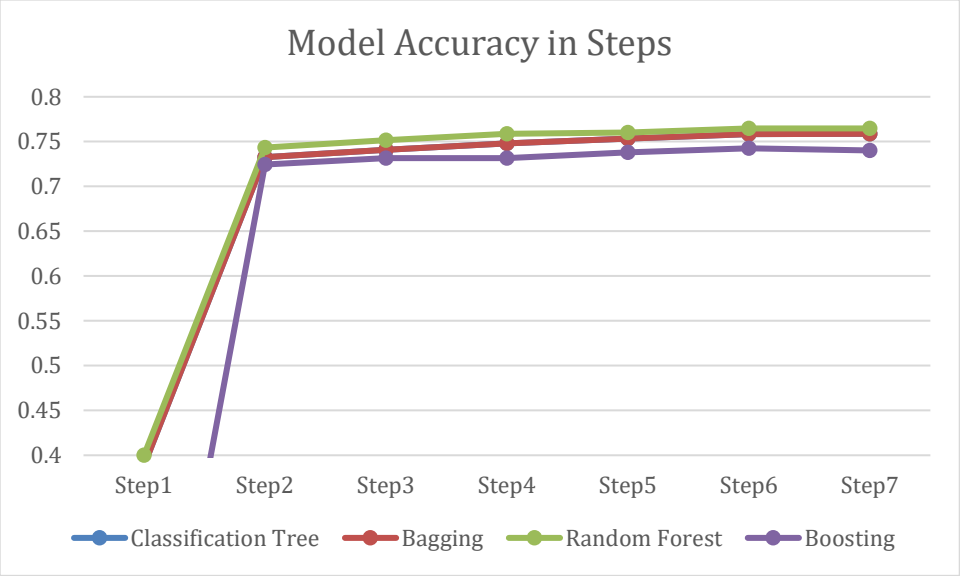

Exhibit 9 Mean Decrease Accuracy

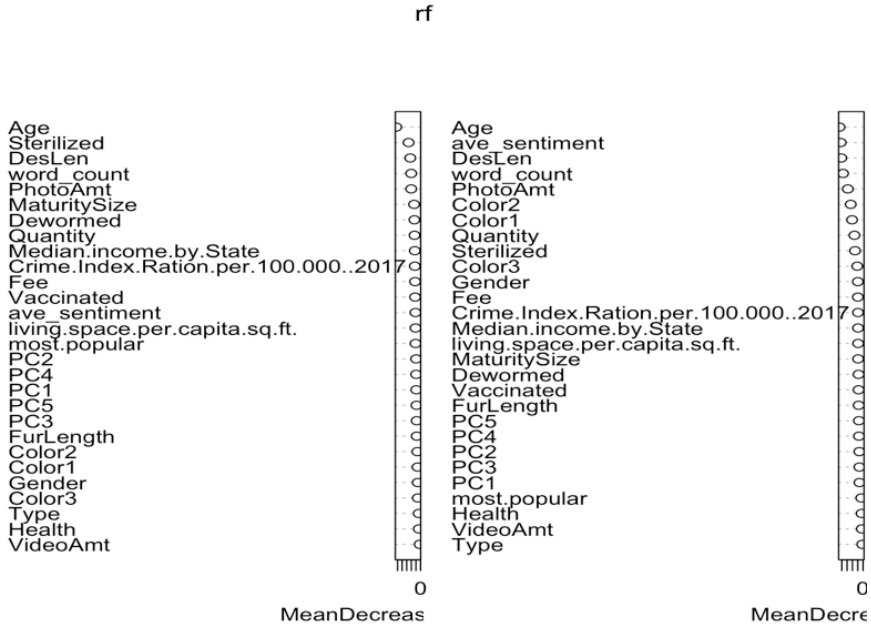

Exhibit 10 Neural Network Plot

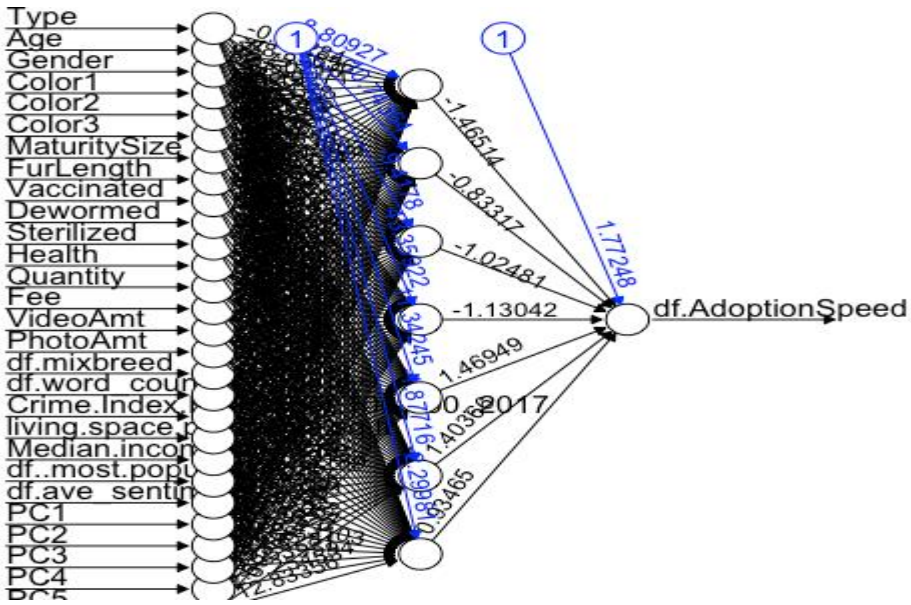

Exhibit 11 Naive Bayes Accuracies (Original v.s. Binned categorical variables)

	Step1	Step2	Step3	Step4	Step5	Step6	Step7
Accuracy	0.3309	0.7085	0.7049	0.7069	0.7091	0.6891	0.7082
Accuracy(Binned)	0.3407	0.7178	0.7147	0.714	0.7147	0.6731	0.7002

Exhibit 12 Accuracy Summary Table

Accuracy	Logistic Regression	KNN	Naive Bayes	Neural Network	Classification Tree	Bagging	Random Forest	Boosting
Step1	NAN	0.3269	0.3407	0.2326	0.3598	0.3885	0.4000	NAN
Step2	0.7316	0.8323	0.7187	0.7360	0.7213	0.7325	0.7432	0.7243
Step3	0.7352	0.8225	0.7147	0.7405	0.7213	0.7407	0.7514	0.7316
Step4	0.7398	0.8185	0.7140	0.7363	0.7213	0.7481	0.7587	0.7316
Step5	0.7411	0.8185	0.7147	0.7358	0.7213	0.7532	0.7601	0.7380
Step6	0.7459	0.8083	0.6731	0.7409	0.7213	0.7583	0.7647	0.7425
Step7	0.7430	0.8147	0.7002	0.7403	0.7213	0.7587	0.7647	0.7400

Exhibit 13 The model with highest accuracy (KNN, in Step 2, when k=5)

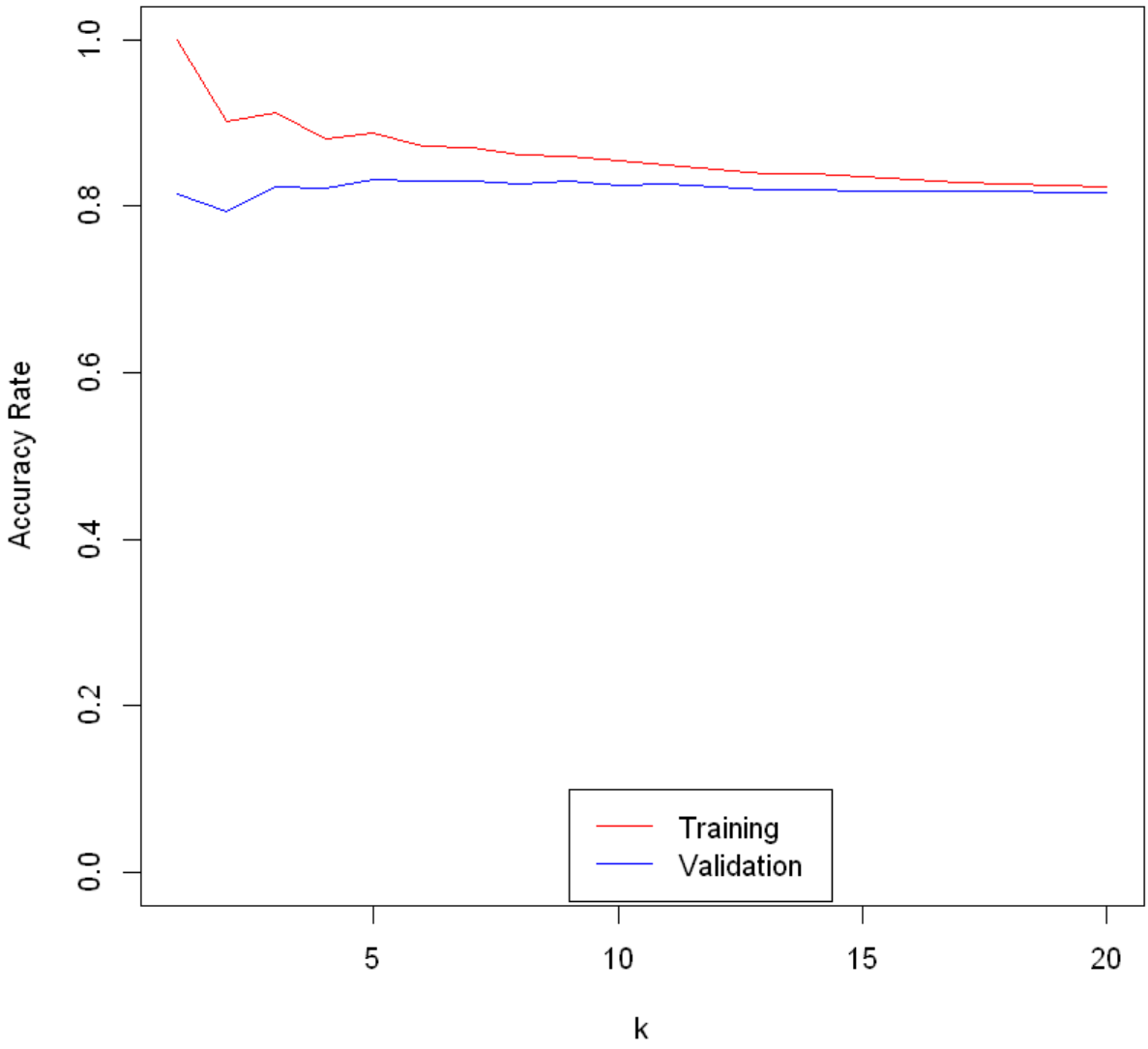

Exhibit 14

prediction	0	1
0	7815	971
1	202	1505

Exhibit 15


adopt
love
will
pleas
home
kitten
puppi
cat
dog
can
good
look
give
month
contact
need
take
home.
care
play
found
call
interest
rescu
like
one
new
owner
friend
vaccin
now
time
hope
neuter
famili
adoption.
make
male
abandon
litter
food
know
activ
cage
kind
also
dog.
black
vet
just
two
find
help
still
ador
hous
week
pet
someon
spay
back
must
alreadi
around
eat
train
cute
serious
fee
thank
sweet
realli
peopl
forev
stay
old
littl
femal
keep
year
provid
sinc
well
foster
anyon
want
stray
come
healthi
get
old.
let
mother
pls
long
live
current
mix
white
feed