Data science, prediction, and forecasting exam

Introduction

Overpopulation in animal shelters has been an issue in the USA for many years now. There are simply too many animals without homes and too few shelters to house, help, and rehome them. This overpopulation leads to millions of cats and dogs being euthanized every year (Kass et al., 2001; Spehar & Wolf, 2020). Much research and work has been done to better understand why this is and how to help with this crisis. This paper explores a theoretical way to help the shelters allocate resources to maximize adoption rates using a supervised machine learning model made with the package XGBoost in R.

In a survey that asked adopters why they adopted their pet out of all the ones in the shelter people responded that talking to the shelter staff about the animal, talking about the behaviour and health of the animal had a much bigger effect on their decision than just reading the information on the cage cards. Getting to interact with the animal and not just see it in its cage was another important factor in their decision (Weiss et al., 2012). For cats, the complexity of the environment they were being kept in, if they had toys, and often being handled increased their chance for adoption (Fantuzzi et al., 2010; Gourkow, 2001). Similar results have been found with dogs, alongside house training them increasing adoption rates (Luescher & Tyson Medlock, 2009; Protopopova et al., 2012).

The problem with solutions like those above this is that it takes time and resources to implement which shelters do not tend to have. This is where the theoretical use case of the models made in this paper could come in. If the shelters could focus their resources on the animals that are predicted to have a longer stay it could shorten their time spent there.

What determines an animal's adoptability? There has been a decent amount of research on the topic, the articles cited in this paper will primarily be from the USA as the data used in the analysis is from there, however, a few studies will be from the UK. It seems that regardless of animal type age is always a determining factor, older animals were significantly more likely to get euthanized than younger or middle-aged animals (Brown et al., 2013; Brown & Morgan, 2015; DeLeeuw, 2010; Kass et al., 2001; Kopelman et al., 2002; Lepper et al., 2002; Svoboda & Hoffman, 2015). The sex of the animal is also a consistent factor for adoption rates in both cats and dogs (Brown & Morgan, 2015; DeLeeuw, 2010; Kopelman et al., 2002; Lepper et al., 2002). An often-mentioned factor for dogs is the breed, breeds in the category "Bully breeds" like bulldogs were also less likely to get adopted than other dog breeds (Brown et al., 2013; Kopelman et al., 2002; Lepper et al., 2002; Svoboda & Hoffman, 2015). In relation to breed, the purebred status had a significant effect on adoptions, studies

found that purebred dogs got adopted more often than non-purebreds (DeLeeuw, 2010; Kopelman et al., 2002; Lepper et al., 2002). The purebred status will not be part of the analysis as the data does not include that information directly, indirectly the breed names do include if they are a mix or not, so the model still has some information on the purebred status of the dog.

Whether fur colour affects adoption rates is a debated topic. It is commonly believed that animals with dark or black fur colour have a disadvantage when it comes to getting adopted. This "phenomenon" is often called "black dog syndrome" sometimes more specifically "Big black dog syndrome", despite the name it does also affect cats (CBS Pittsburgh, 2012; Nakano, 2008). Despite how many shelter workers believe this phenomenon affects the adoption rates it has not consistently been found in shelters, with some studies finding no significant effect when it comes to adoption (Brown et al., 2013; Brown & Morgan, 2015; Svoboda & Hoffman, 2015; Woodward et al., 2012). However, one study on dogs and one on cats found "not having a primarily black coat" to be a predictor for higher levels of adoption (DeLeeuw, 2010; Kogan et al., 2013) and another listed colour as a factor in relation to adoption (Lepper et al., 2002), so while having a black coloured fur specifically might not affect adoption, there is some evidence that colour in general does.

Other relevant factors that were not found in every paper were, whether or not the animal is injured (Kopelman et al., 2002; Lepper et al., 2002) and the "intake type", that is if the animal was found as a stray or given to the shelter by a previous owner (DeLeeuw, 2010; Lepper et al., 2002). Only one study mentioned neuter status and the shelter itself as a factor (Kopelman et al., 2002)

The data used in this analysis is from Kaggle, where it was put up by Aaron Schlegel in 2018 (Schlegel, 2018). The data was originally made public by the city of Austin, Texas and is continually updated (Austin Texas, 2023). The data is collected from the Austin Animal Center from 2013 to 2018. It contains information on the animals that come through the shelter, both at the time the animal entered the shelter (intake) and then how it left, either via adoption or other means (outcomes). For a breakdown of the data used in this analysis please see table 2. in the appendix.

Methods

The analysis will be made using RStudio and the XGBoost package. The code used can be found at this public GitHub: https://github.com/au636396/Data-science-exam-Cecilie-Vestergaaard

Preparation of the data:

The dataset needs some tweaks to fit our needs, on its own it contains some amount of irrelevant information. While this dataset contains 4 categories of animals (see appendix), the Birds and the Others have been removed as the general research on the topic primarily covers cats and dogs, 339 birds were removed, and 4428 others were removed. The animals that were tagged as intake type "Euthanasia Request" have been removed as they don't enter the shelter with the assumption of adoption, 240 animals were in that category. The sex column includes both the sex of the animal and

not whether or not the animal has been spayed or neutered (see Table 2. In appx.). Since the research mentions sex but not the other as a factor for adoption times the column will be split into two. One that just contains the sex of the animal and one that contains info on spayed or neutered status, renamed fixed or not_fixed.

In cases where an animal has more than one colour listed the most prominent one will be considered its only colour; this brought the number of unique values down from 500 to 57. Some further renaming was done of those categories that only had one or very few entries to bring the number of colours down to 52. This scale-down in the number of colours was done in an attempt to make the models have an easier time finding patterns in the data, if many of the colours only had one entry they would have little to no effect on the predictions the models would make. The same method of removing the sub-names was performed on the breed column, bringing the number of unique breeds down from 1997 to 426. Unlike colour, the breeds are too different and there are too many to nicely rename them, so no more cleaning was done on that column.

The outcome of the models would be the amount of time spent in the shelter, the data has that measured in days, with 471 unique amounts of time ranging from 0 to 1606 days. These many possible outcomes would make it almost impossible for the model to predict the right one. Therefore, a week and month column was added, bringing the number of outcomes down to 119 and 43 respectively. In the months column a big number of months were only populated by one datapoint, this issue was smaller in the weeks measure. Therefore, the week measure was the one chosen for the analysis.

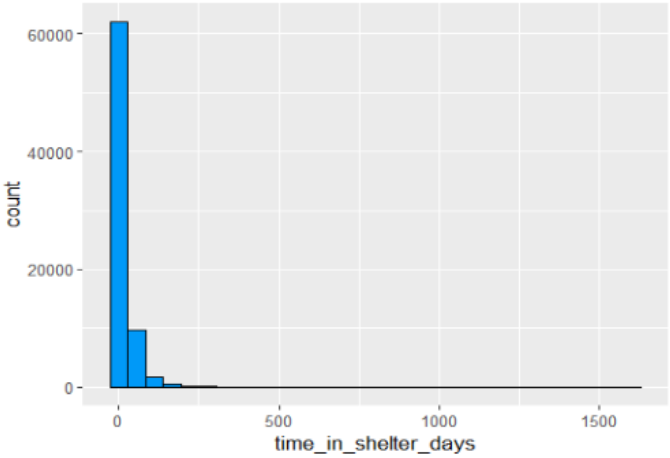

Figure 1. Histogram of time spend in the shelter.

One major problem with the data is the lack of data on animals that stay for longer periods of time, as seen in Figure 1. most of the data points lay close to the low numbers, with almost no data for animals that stay for longer periods of time, the median amount of time is 5 days. In the shelter where this data has been collected it is simply rare for an animal to stay for longer than a month or so, therefore there is very little data on the longer-staying animals. I suspect one of the reasons for this is that this shelter often transfers animals to other shelters, so some of the animals have just been moved to another shelter and not adopted in that short time. This is of course a good thing for this shelter but not so for our analysis. There are multiple ways of getting around this, one could remove some of the data points below 5, recategorize the data, either as a binary that just determines if an animal will have a short or a long stay, or just recategorize data above e.g., 50 days as one category denoting a long stay. In this

paper we'll attempt both an outcome measure of 3 values or a "trinary" and making the dataset more even by reducing the number of data points below 5 days. These two methods should both result in slightly more balanced data in different ways. The 3 values outcomes will represent the animal staying: "one month", "over one month", or "under one month", (In the markdown, they are coded as, 1, 2, and 3) where the over one-month value includes all the animals that stayed for longer than a month. This setup was chosen as it gives a little more information to the shelters than e.g., a binary outcome would, they will either know if an animal is in for a short, medium, or long stay and then allocate resources accordingly. For the "even" dataset 90% of the values below 5 have been removed, this makes the dataset a lot smaller, going from 74.659 observations to 21.233 observations. The "even" in the "even" dataset is in quotation marks as the dataset still has a majority of lower values, however, it is significantly more even than the normal dataset.

Models

The code used in the making of these models is mainly from two tutorials from Dale Kube and Savvy Sahai (Kube, 2019; Sahai, 2022). The models will be made with the XGBoost (Extreme Gradient Boosting) package, as it is quite popular for supervised machine learning tasks. This is partly because it offers many useful build-in features. It has Parallel computing and as a default will use as many threads as the computer has cores, this makes the training much faster. It offers regulation that reduces overfitting. It comes with other nice features such as a build in cross-validation function for tuning the model, the models are easily saved and loaded, and it handles NAs well. More specifically the XGBoost package was used to make a decision tree model. Decision trees use branching paths to categorise items, in the case of gbtree it grows one tree after another learning updated weights from the mistakes the prior tree might have made. An example of a tree an XGBoost model made can be found in Figure 2.. XGBoost decision tree models work a little differently to most others concerning how tree pruning is performed . Tree pruning means removing noncritical branches from the tree making the model more efficient and reducing the size. This process is generally done while the tree is "growing", whereas in XGBoost it is done after the tree is fully grown.

Based on the research mentioned above 3 models with different input variables have been made, starting with a simple model including all the factors all papers could decide on (model 1), then one also including colour as a fair number of papers mention it being a factor (model 2), and finally one including things mention by a minimum of two papers (model 3).

Model 1: Animal type, age, breed

Model 2: Animal type, age, breed, coat colour

Model 3: Animal type, age, breed, coat colour, intake type, intake condition

The models were tuned with the help of the XGBoost documentation (xgboost developers, n.d.) and a tutorial by Manish Saraswat (Saraswat, 2022). Multiple boosters, objectives, and evaluation metrics fitting the nature of a categorisation on many categories were attempted and the best were chosen based on the accuracy of the model when tested against the test dataset and the build in cross-validation function in XGBoost. The same was done on the learning rate(eta), the maximum depth of trees, gamma, subsample, nrounds, and alpha values. The Booster chosen was gbtree, as it gave the best accuracy with our data and decision trees are good at this type of classification task. The objective is set to multi:softprob, this will do a multiclass classification using the softmax function, which outputs the probability that each datapoint belongs to each class (xgboost developers, n.d.). The evaluation metric was mlogloss, which uses the log loss function to measure error in multiclass classification.

Figure 2. Example of a decision tree. This is tree nr 15 from model 1 trained on the "even" dataset.

As the accuracy of all the model's predictions of time in weeks is expected to be quite poor and to see whether different outcomes would produce higher accuracy, the models will be run again with the 3 values outcome and with the "even" dataset. In the case of the 3-value outcome, this is not quite what these models were supposed to predict, as it does not give as much information as the number of weeks would, however, it would be interesting to see if it improves the accuracy.

All in all, this comes out to 9 different models that this paper will run, the 3 different models from above all run once with the standard data, the one with a 3 values outcome, and the "even" data.

Results

	Standard	Time measured in 3 outcomes	"Even" dataset
Model 1	44.90%	76.53%	69.18%
Model 2	44.03%	76.45%	68.74%
Model 3	48.53%	76.47%	69.06%

Table 1. The accuracy of the 3 models in the standard data, the 3 values outcome, and the "even" data.

Discussion

The results for the standard models that try and output the number of weeks an animal will stay are quite poor, performing with around an accuracy of 45%. Out of the three models, the best performing is model 3 with an accuracy of 48,53%, this is unexpected as I would have thought the added complexity would confuse the model, however, that was not the case. There is not much of a difference between models 1 and 2, this would indicate that colour, at least in this shelter, does not have an effect on time spent in the shelter. They are all performing much better than a random sort would, indicating that there are indeed trends in what animals get adopted faster than others. One major reason the models perform so poorly is that the models are only predicting low values, e.g., 1 and 2, corresponding to 0 and 1 weeks spent in the shelter. As mentioned above I suspect this is because the outcome data is so skewed towards the lower values.

This is where the simplified outcome comes in, the models did much better with the simple outcome variable, performing with around an accuracy of 76%. In this setup, random guessing would also be easier, with a 33,33% accuracy. The best performing model is now technically the first, however, the difference between them is negligible. While 76% is a big improvement over 45%, it would still be too unreliable to use in a shelter.

To investigate whether the uneven distribution of the number of weeks spent in the shelter was a reason for the not great results in the first rounds of models, they were all trained again on the "even" dataset. All the models performed with an accuracy close to 69%. This means that the model's accuracy improved with 24% on average, by training with a data set that was not as skewed.

To investigate whether the combination of the trinary outcome and the "even" dataset would further improve the accuracy, model 3 was run with both of them. Interestingly the accuracy for this combination is only 75.95%, about the same as in the trinary outcome models, if not a little worse. So, while reducing the outcome to 3 values and reducing the amount of data points below 5 both

improved the model accuracy by a lot, the combination of the two does not improve the accuracy more than just the level the models performed at when using just the trinary outcome.
All in all, the models do a decent job of learning the predictive outcomes of how long an animal is likely to stay in a shelter. The unevenness inherent in this data makes the learning process harder for them and when it is somewhat controlled for, they do a much better job. Do these models perform well enough to be useful in a shelter? No, they are sadly not consistent enough to be useful in a real shelter, the best model would still be wrong about 24% of the time. It is way too unreliable to put in a shelter to help allocate resources. It is possible that other shelters with their own data might be able to produce better models, especially if the people adopting in that areas are consistent in what they prefer, in terms of age, breed, etc..

A thing to keep in mind when talking about the possible values of such models is that the subject matter is living animals and potentially how they would be treated. If these models were implemented into the workflow at a shelter it would mean that some animals will not get a lot of resources purely because machine models predicted they would not stay for very long. This could in theory mean that a puppy will not be interacted with as much as some other dogs, as the resources to do so are going to other dogs. While this might still result in the best overall outcome for the shelter as it will minimise the number of animals that stay longer, it might then make some prior shorter stays longer as a result. This is a very utilitarian way to allocate resources and that could easily be seen as inhuman or "cold". So even if these models performed at 100% accuracy and were used purely to try and shorten longer stays, there would still be ethical concerns. It is also unrealistic to imagine an omniscient model that never makes mistakes, it is unavoidable and begs the question of whether models such as this is worth the possible mistakes and in what situation would it be worth it.
While these models are intended for shelters to help make decisions on allocating resources to give all the animals an equal opportunity to get adopted, they could easily be used in other ways. If a traditional shelter or so-called "kill shelter" (Puyear, n.d.) used models like these they might want to save resources by euthanising the animals with a predicted long stay. This would inevitably end in animals that would have been adopted never getting the chance. However, one could argue that that already happens, therefore, whether or not it is decided by a model or a human it is irrelevant. It might even help alleviate the mental strain making decisions like that has on shelter workers (Andrukonis et al., 2020; Reeve et al., 2005).
Another issue with implementing models like this is that they inherently carry the bias from the data into the future, even if they are updated regularly. This effect is clearly shown in examples of using machine learning to "optimally" position patrol police presence. Using collected data on crime rates, their location, and the time of day these models are then trained to predict what areas and at what times crime will occur, smart right? In theory, this will allocate the "resources", that is police officers, more effectively, catching and hopefully preventing crimes before they happen. However, that is not

what happened in the areas this was used. The crime rates stayed high in high-risk areas (Reese, 2022). This is because the police being there now catch more criminals as that is where they are looking, artificially inflating the numbers higher than other areas. This new data is then fed into the algorithm, creating a feedback loop where everything will forever be based on the initial data the model was trained on. The same fear comes into play with these models, hopefully to a lesser extent when used as intended. But there is some concern that whatever bias is present in the data now will stay the same even if newly collected data will be fed into it. Similarly, the models would all need to be made for each shelter as they would get unreliable if one shelter used another's shelters data, as it would inherit a bias that might then "spread" to the new shelter.

Ideally, the shelters around the USA would be given more resources to prevent the potential use of machine learning models being necessary at all. In the case of cats, another less resources heavy solution is to increase the amount of trap-neuter-return programs, which involves neutering stray cats and returning them to the wild, in the long run, this decreases the number of stray cats while not euthanizing any of them (Levy et al., 2014; Spehar & Wolf, 2020).

Future research

As the data used in this analysis contains so few longer-staying animals since the shelter it has been collected from is quite good at minimizing the stay of the animals, with a median stay of only 5 days long, the models had a hard time learning much about them. Perhaps if the data was collected from a shelter that in general had longer stays the models would have performed better, similarly more data on longer stays would have improved the models.

Another possible way to make this a useful tool would be to train a model on the outcome of an animal instead of the time spent in the shelter. The outcome ranges from adoption to euthanasia. This paper did not make any such models as the ethical implications of an algorithm predicting the death of an animal are even more heavy and complex than the models made here.

Conclusion

While the models made in this paper would not be good enough to place in real-world shelters, it does show the potential for models with higher accuracy. They could be put into animal shelters to help allocate resources, ensuring the best outcome for the most animals. However, doing so is not without ethical implications. Models like those could easily be misused and end in regrettable outcomes, especially since the models are going to be wrong at times.

References

Andrukonis, A., Hall, N. J., & Protopopova, A. (2020). The Impact of Caring and Killing on Physiological and Psychometric Measures of Stress in Animal Shelter Employees: A Pilot Study. *International Journal of Environmental Research and Public Health*, *17*(24), Article 24. https://doi.org/10.3390/ijerph17249196

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*(1), 2–18. https://doi.org/10.1080/10888705.2013.740967

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. https://doi.org/10.1080/10888705.2014.971156

CBS Pittsburgh. (2012, August 15). *"Black Dog Syndrome" Affecting Adoption Rates At Shelters*. https://www.cbsnews.com/pittsburgh/news/black-dog-syndrome-affecting-adoption-rates-at-shelters/

DeLeeuw, J. L. (2010). *Animal shelter dogs: Factors predicting adoption versus euthanasia*. https://soar.wichita.edu/handle/10057/3647

Fantuzzi, J. M., Miller, K. A., & Weiss, E. (2010). Factors Relevant to Adoption of Cats in an Animal Shelter. *Journal of Applied Animal Welfare Science*, *13*(2), 174–179. https://doi.org/10.1080/10888700903583467

Gourkow, N. (2001). *Factors affecting the welfare and adoption rate of cats in an animal shelter* [University of British Columbia]. https://doi.org/10.14288/1.0090034

Kass, P. H., New, J. C., Scarlett, J. M., & Salman, M. D. (2001). Understanding Animal Companion Surplus in the United States: Relinquishment of Nonadoptables to Animal Shelters for Euthanasia. *Journal of Applied Animal Welfare Science*, *4*(4), 237–248. https://doi.org/10.1207/S15327604JAWS0404_01

Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2013). Cats in Animal Shelters: Exploring the Common Perception that Black Cats Take Longer to Adopt. *The Open Veterinary Science Journal*, *7*(1). https://benthamopen.com/ABSTRACT/TOVSJ-7-18

Kopelman, S., Weber, J. M., & Messick, D. M. (2002). Factors influencing cooperation in commons dilemmas: A review of experimental psychology research. In National Research Council (Ed). The Drama of the Commons. *Washington D. C.: The National Academies Press*, 113–156.

Kube, D. (2019, February 26). *RPubs—XGBoost Iris Classification Example in R*. Rpubs. https://rpubs.com/dalekube/XGBoost-Iris-Classification-Example-in-R

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*(1), 29–42. https://doi.org/10.1207/S15327604JAWS0501_3

Levy, J. K., Isaza, N. M., & Scott, K. C. (2014). Effect of high-impact targeted trap-neuter-return and adoption of community cats on cat intake to a shelter. *The Veterinary Journal*, *201*(3), 269–274. https://doi.org/10.1016/j.tvjl.2014.05.001

Luescher, A. U., & Tyson Medlock, R. (2009). The effects of training and environmental alterations on adoption success of shelter dogs. *Applied Animal Behaviour Science*, *117*(1), 63–68. https://doi.org/10.1016/j.applanim.2008.11.001

Nakano, C. (2008, December 6). *Black dog bias?* Los Angeles Times. https://www.latimes.com/style/la-hm-black6-2008dec06-story.html

Protopopova, A., Gilmour, A. J., Weiss, R. H., Shen, J. Y., & Wynne, C. D. L. (2012). The effects of social training and other factors on adoption success of shelter dogs. *Applied Animal Behaviour Science*, *142*(1), 61–68. https://doi.org/10.1016/j.applanim.2012.09.009

Puyear. (n.d.). *HUMANE SOCIETY: Exploring differences between "no-kill" and traditional shelters*. The Daytona Beach News-Journal. Retrieved June 1, 2023, from https://www.news-journalonline.com/story/news/local/flagler/2018/07/04/humane-society-exploring-differences-between-no-kill-and-traditional-shelters/11601053007/

Reese, C. (2022, February 23). *What Happens When Police Use AI to Predict and Prevent Crime?* JSTOR Daily. https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/

Reeve, C. L., Rogelberg, S. G., Spitzmüller, C., & Digiacomo, N. (2005). The Caring-Killing Paradox: Euthanasia-Related Strain Among Animal-Shelter Workers1. *Journal of Applied Social Psychology*, *35*(1), 119–143. https://doi.org/10.1111/j.1559-1816.2005.tb02096.x

Sahai, S. (2022, December 20). *How to apply xgboost for classification in R* -. ProjectPro. https://www.projectpro.io/recipes/apply-xgboost-for-classification-r

Saraswat, M. (2022). *Beginners Tutorial on XGBoost and Parameter Tuning in R Tutorials & Notes | Machine Learning*. HackerEarth. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/

Spehar, D. D., & Wolf, P. J. (2020). The Impact of Return-to-Field and Targeted Trap-Neuter-Return on Feline Intake and Euthanasia at a Municipal Animal Shelter in Jefferson County, Kentucky. *Animals*, *10*(8), Article 8. https://doi.org/10.3390/ani10081395

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*(4), 497–506. https://doi.org/10.7120/09627286.24.4.497

Weiss, E., Miller, K., Mohan-Gibbons, H., & Vela, C. (2012). Why Did You Choose This Pet?: Adopters and Pet Selection Preferences in Five Animal Shelters in the United States. *Animals*, *2*(2), Article 2. https://doi.org/10.3390/ani2020144

Woodward, L., Milliken, J., & Humy, S. (2012). Give a Dog a Bad Name and Hang Him: Evaluating Big, Black Dog Syndrome. *Society & Animals*, *20*(3), 236–253. https://doi.org/10.1163/15685306-12341236

xgboost developers. (n.d.). *XGBoost Parameters—Xgboost 1.7.5 documentation*. Retrieved May 27, 2023, from https://xgboost.readthedocs.io/en/stable/parameter.html#parameters-for-tree-booster

Appendix

Columns:	Values	Further information
animal_id_intake	1 per animal	IDs for the animals.
intake_type	5 factors, "Stray", "Public Assist", "Owner Surrender", "Euthanasia Request"	The circumstance the shelter revised the animal in. “Euthanasia Request" was removed for this analysis.
intake_condition	8 factors, "Normal", "Injured", "Aged", "Sick", "Other", "Feral", "Pregnant", "Nursing"	The health condition the shelter revised the animal in.
time_in_shelter_days	Range from 0 to 1606	The amount of time in days the animal spent in the shelter.
animal_type	4 factors, Bird, Cat, Dog, Other	The type of animal. Only cat and dog were kept for analysis.
sex_upon_intake	5 factors, "Neutered Male", "Spayed Female", "Intact Female", "Intact Male", "Unknown"	The sex of the animal and whether it has been sterilised.
age_upon_intake_(years)	Range from 0 to 25	The age of the animal when it entered the shelter.
age_upon_intake_age_group	10 age groups, in increments of 2,5	The age of the animal when it entered the shelter divided up into 10 ranges.
breed	1997 factors, e.g., ‘Labrador Retriever/German Shepherd’	The breed of the animal. For the analysis only the first breed listed will be used.
color	500 factors, e.g., ‘Black/White’	The colour of the animal. For the analysis only the first colour listed will be used.

Table 2. overview of the data