**Final Report - Pet Adoption Model**

Rodolfo I Cardenas

Math 111A, UC San Diego

***Introduction***

Currently, animal shelters have received an increase amount of adoptions due to COVID-19. When you try to adopt a pet you go in person to an animal shelter or look online for all the pets that they currently have available. After doing this first step you choose one that meets your own standards. Some questions are asked by the animal shelter and then on you go with your new pet. What happens to the pets that do not meet your standards? Sometimes other persons will have different preference from you and end up adopting them but what if they do not? Some dogs are not attractive to today’s standards or could have characteristics that makes them hard to adopt. A senior dog will tend to have more health problems than a younger dog which will lead it to be less viable for someone that does not have the resources to sustain its health problems. Older dogs could also not be used to another animals therefore will need to have training to interact with other pets, something that cannot be done by a normal person. (1)

Letting people choose the pet that they prefer will always lead to some pets being left aside as there are trends, preferences in size, color and breed of pets. This process eventually leads to a pet not getting adopted or taking a long amount of time to get adopted, this could lead to depression in the pet that could lead to trauma and even euthanizing for some animal shelters if it exceeds the maximum amount of days that the pet is allowed to stay in the shelter (This doesn't apply to all animal shelters).

There are patterns when people adopt a pet, preference in smaller dogs than larger dogs or a certain color instead of a mixed color pet. (2) Pets chosen are not random as it is not an easy decision to adopt a pet therefore people want to pick the best match for them. “Characterizing Unsuccessful animal adoptions” in Nature.com is a model where they calculate the risk of return after the adoption of the pet, as we know there are persons that believe are ready and have all factors in favor to adopt but something could happen which leads them to returning the pet. The model uses categorical values as “too active”, “landlord issues”, “health of animal”, etc. (3) This type of model takes the use of Logistic Regression to quantify these variables and therefore create predictions that can lead to an approximation of what could happened to a dog. This model takes the same process but applied into the step before the adoption happens.

A model for dog abandonment that exists is “Human–dog relationships during the COVID-19 pandemic” published in natura.com they used a digital questionnaire to take account of how COVID-19 affected the life of the respondents which was used to calculate the quality of life of the dog that they adopted. (4) According to their results, the characteristics of the owner that adopted the dog have a high influence on the quality of life of the dog, which brings us back to trends that people have could lead to dogs being discriminated for its characteristics.

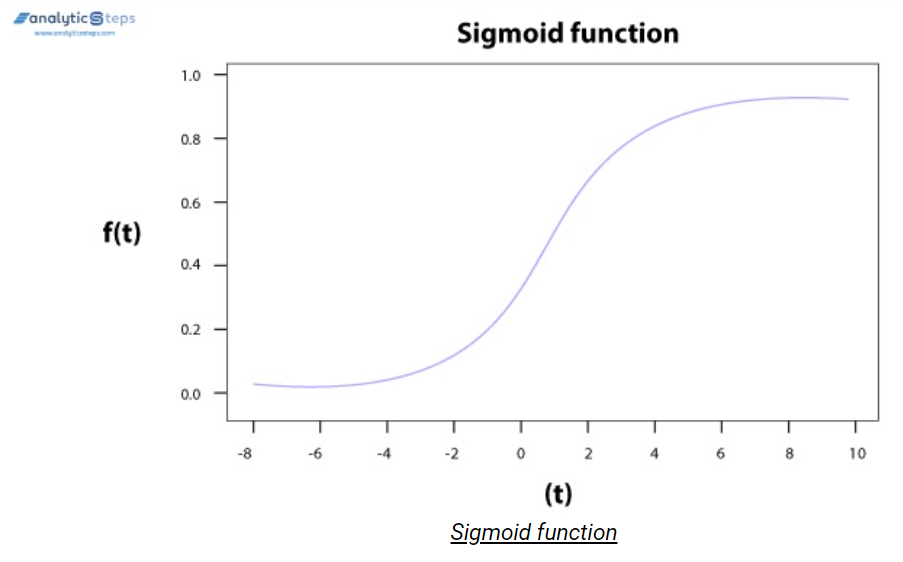
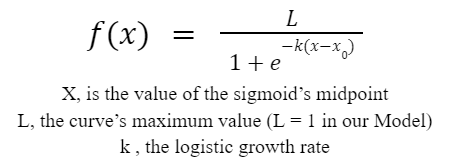
This model marks the probability of a dog on being adopted which will highlight the importance that we need to give to those who have the highest percentage to the point of giving incentives or more publicity to help them get adopted and hence the animal shelter could be prepared for the pet to start for a longer amount of days. The question for this model is “Could we build a model to predict the likelihood of a pet being adopted?”

***Model Description:***

In this mathematical model we take the use of logistical regression. (7) Which can be related to linear regression as this one goal is to create a fit for a straight line with continuous data points, where Beta is the slope of the line. In comparison, logistic regression does a similar thing where we fit all our data into this function and we will take the use of machine learning with python to solve for our betas that could be represented as our slopes to find our own graphical representation for this model.

Our model includes categorical values that count as a success or a failure of our goal, pet adoption in animal shelters. If the outcome of a pet is not adopted, we will consider that as a failure and if the pet is adopted we will consider that a success. In order to take use of this data with the logistical regression method we took all outcomes of our data and replaced them with a 0 if it was a failure and a 1 if it was a success. (8)

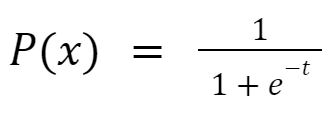
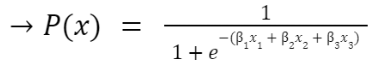
The graph of a logistical regression model will create a standard sigmoid function which is also known as an S-graph as we have our bottom asymptote at 0 and our top asymptote at 1 because we only have two outcomes, 1 and 0. (9) The logistic function is the following:



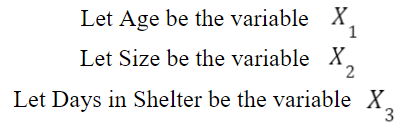
Using this function for our model we can see that L will be replaced with only 1 as that is the maximum value for our curve. We also want to apply multiple independent variables (Age, Size, Days in Shelter) to our function. In order to do this, we need to replace our t with a linear combination that is made with x’s and betas.



Applying this change into the logistic function we have:

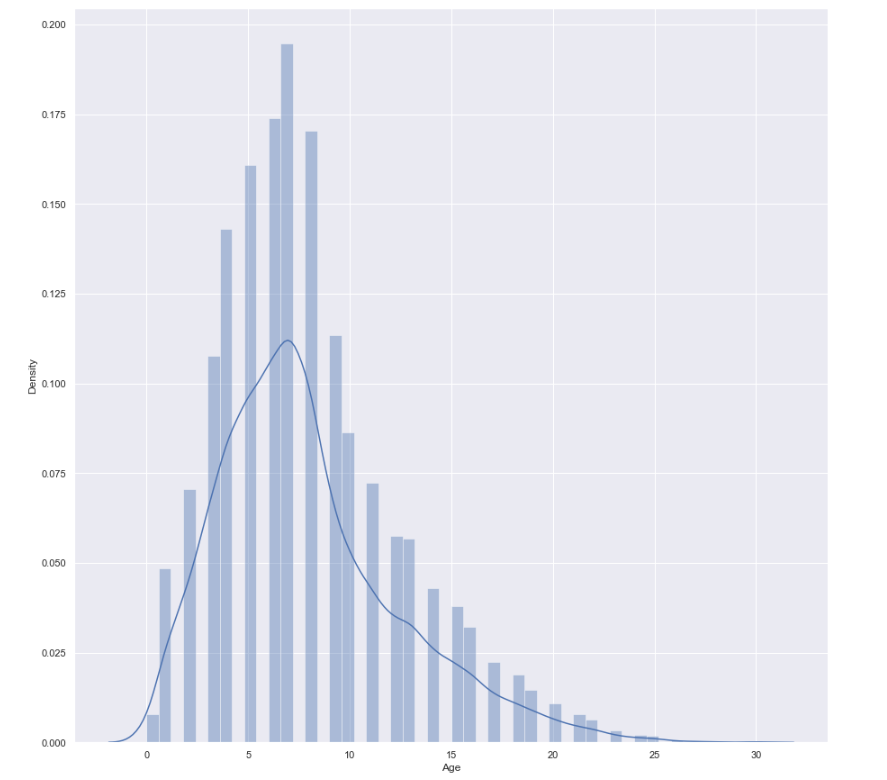
P(X) will give us the probability of a dog being adopted which makes this our dependent variable.



***Data Description:***

We used a data frame provided by the County of Sonoma Department of Health Services. (5) This data frame was particularly picked as it is the closest to San Diego County in which we could find current and updated data available for free online. The data frame originally included 28 columns with a lot of useful information about animal adoption in Sonoma county for which we only used: Type, Size, Date of Birth, Outcome Date, Days in Shelter, Outcome Type. The last time that the data frame that we used was updated is November 23, 2021. In this data frame we have information available for multiple types of pets including dogs, cats, birds, etc. In our model we only took use of Cats and Dogs as these two were the ones with the most data therefore we have enough to create a prediction about which pets will be adopted and have a probability chance for each pet.

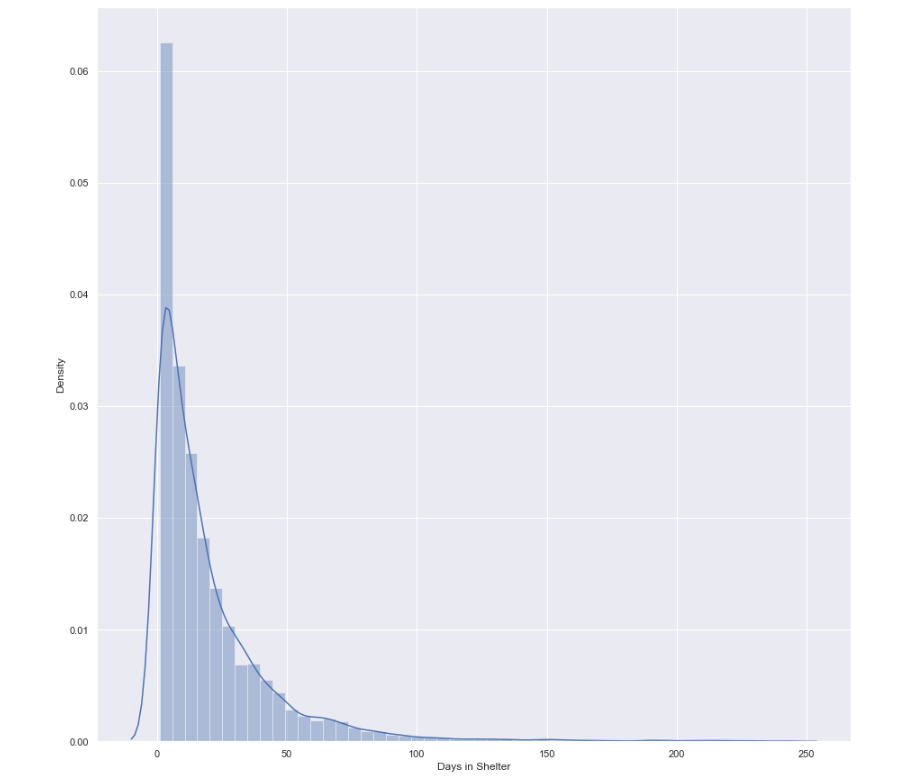
One of the variables that were considered is Age. Here is where we make our first assumption for this model, we will assume that a dog of 11 or more years of age is considered a senior dog. Unfortunately, in the data frame that we are using we do not have a specific variable for age, but in our case we used the date of birth which is included as a column and also the outcome date which is also included, this date tells us when the end result of the pet happened, if it was adopted, euthanized, returned to owner, etc. By taking the amount of days between both dates we can divide it by 365 and get an estimate of years that the pet has of age.



We can notice that most of the values in age are between [0,20] with a max close to 7.5 years that is where most of our data is concentrated.

Size of the pet is the second variable that was taken into account. In this case we were given categorical values for which each pet size was determined, for example instead of being given a specific amount of pounds that a pet was weight at, we only had available estimates: Kitten, Puppy, Toy, Small, Med, Large and X-Lrg. (6) For this calculation we estimated the size that each categorical value should have to approximate the specific weight, Why? Because to take use of a logistical regression model applied in python we will need to make any categorical value numerical.

Days in the shelter is something that is very relevant to the probability of a pet being adopted because as more days pass on the shelter the pet will have more chances each day of being adopted, Why? Because each day can be seen as a new event, new possibility of a person walking in and being interested on adopting them.



In this graph we can notice that most of data for days in the shelter is concentrated in the first part of the graph, around 0 and 10 days and this is because the column in our data frame that includes the failures where pets are not adopted do also include pets that were returned to their owners therefore this explain the peak on a short amount of days in the shelter for most of the pets that pass thru the animal shelter.

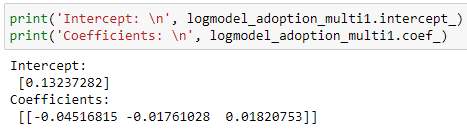
To find this probability we will be considering 3 factors: Age, Size and Days in the shelter. The hypothesis is that young smaller pets with more days in the shelter will have a higher probability of being adopted compared to senior larger pets with less days in the shelter.

***Model Analysis:***

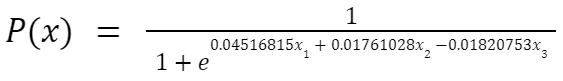
We used python sklearn features to create a logistic regression model that can make predictions with the data frame that we are using by taking use of a process of machine learning. We first take our Data Frame and we split it into 70-30 proportions. 70% will be used to train and build our model and 30% will be used to test the model that we created and create a prediction with it. (10) We use 70-30 because this is a common practice in data science, 80-20 was not used as our data frame is not big enough for it. We create a model in order to estimate our betas with the information that we have, we need to find the “slopes” to be multiplied with our independent variables. We create a prediction column with the tested data. In our situation we already have the results and we know which dogs have already been adopted or not adopted, therefore we are able to compare our predictions and see which ones we got correctly and estimate if our model is accurate. Creating this approximation of betas for three variables is not possible to be represented in a simple graph as we would need four dimensions, one for each independent variable and another one for the dependent variable. Here is an example of successes and failures of pets being adopted with also its probability line. We can see that this supports our hypothesis that older pets have a harder time getting adopted as the line goes lower in the y axis as the pet has a higher age.



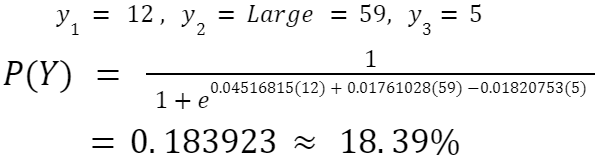
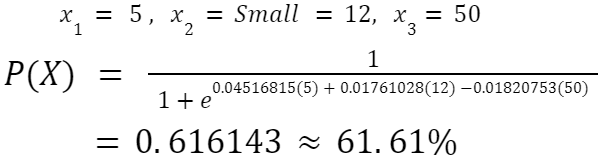
Based on the results from the logistic regression plot, we can see that this supports our hypothesis that older pets have a harder time getting adopted as the line goes lower in the y axis as the pet has a higher age. Following are the regression coefficients results that we got from our data set applying our logistic regression model:



In this case we disregard the intercept as our probability function focuses on the rate of change therefore we want to stay with only the coefficients that apply to each independent variable.  
Applying our results to the starting probability function we get the following equation:



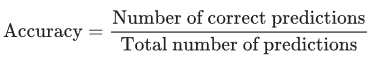
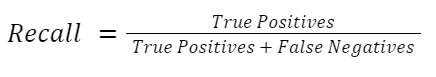
To verify our hypothesis, we have the situation where there is event X and event Y. Let X be the event of having a random 5-year-old small dog which has been 50 days on the shelter. Let Y be the event of having a random 12-year-old large dog which has been 5 days on the shelter.



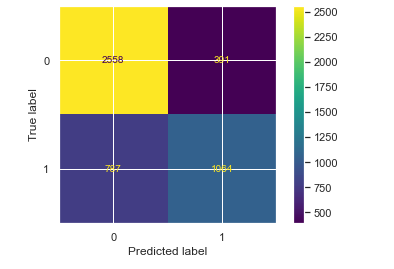
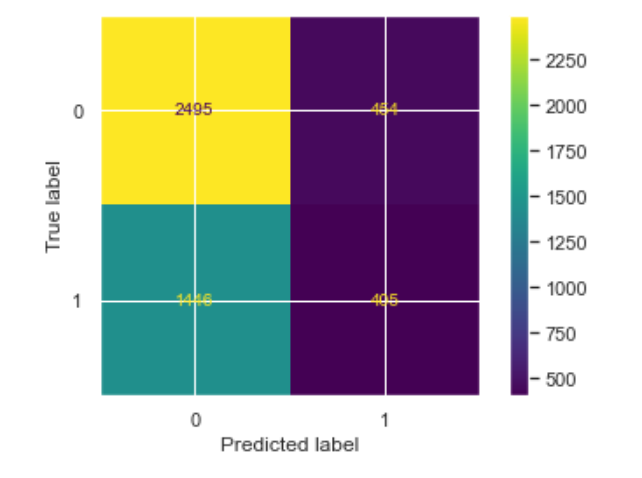
This confirms the hypothesis applying our model as

https://lh6.googleusercontent.com/MkNMSxhjhfb8DmN2TygQNqqWfW5cUjBFQ_JerOiH4FFOOeYcF24E-cNFx0_xm2BmXj8K-En7Qh0H22hHb2pmYOvBsf3oCspouf8jqYakMhSPjLhwmuI9UJoBwZ78VuG7UraBzk6KJn8lhttps://lh4.googleusercontent.com/hUQ59NSO3pbx54XZQY3MmS1zGPL0VntzE084MJkOjGzpasDS3J1bXKmhAW1BbQW_WpvcPqhokxJTqDEnj8AQaj5Vpvf8KX7coG6H_2QIsa4yp-5Hb4GMpGCt0ky9cJ8CDmnH8HuXzjNe https://lh6.googleusercontent.com/tuccY6S5gYg1wBbnwwi00Skuptb9J-qyYHgLDhdRVS_UMLDByNv6OArtqGpnT_M9sYVTgmKOVsQGd0XZ685cp8SKFhSVg5CA0vlPcdpFEmQFXLiKgC8Sy2ROWcqTEQjCnlJseUmsahK2

How accurate is this model? Python has two integrated tools called recall and accuracy. Recall is a way to score how good our estimate was after comparing the original outcome and the predicted one. Accuracy is the fraction of prediction our model got right.



Creating confusion matrixes with python code allow us to have a clear picture of what was compared and how it correctly matched with the original outcomes and also lets us view the proportions to notice if we had an accurate model. In this matrix we will have True positives/negatives and False positives/negatives for the values we can see that 1 represents still only when a pet is adopted and 0 if the pet was not adopted.



One Variable (Age) Three Variables (Age, Size, Days in Shelter)

https://lh4.googleusercontent.com/k7cTVkyHcGLVOhtiKSFS-EQAOOYRrreFJl1bw1ATWqFyvXiFQudmoZ1iHnaruW4AzAyCKfQuZYo04QixGFgAD3x6IwekpdMr7VVmsPvFJdLM5oFag3R8uzj-TJsuV4L271MT-R2IGJhU https://lh3.googleusercontent.com/EtkZEHd7EVUY_uYw3e_CO2_WGZjPPX2jQIEY6b3YiDAOjTTUh3m8w8MeWnjijKLmV4BKslHkAdczH5Wqx7JpYaEWyrjSrQvmJdTs-G-D34PzLRH4lDPIOxq3IxZUTyvd3yfjU9Ld65fC

https://lh5.googleusercontent.com/KH1Jn12xF289qimSrAg8KeFApvEFpOa8Sx4WWHrVILjdikSY6QAbgQNDVAwHg3QzTTzsgmGbzI8HFhhaMfpif3LmBlAL4sd8xlNCunlDU6loNcNMEyq3OT8q4RHly1LlxsrtlS4IUklR https://lh3.googleusercontent.com/AuMk6dZxPN5fhgHE2MocXN0MqHjfxPTQ67JGxvPqe6JIIiiC5pR9EjZs0WeEb4SOg9nYqXhiaWikzAVUbESgwYqil-Z5cVWQgxeQHjuB6mHG534ih4e1NhAqLrB2cDX-t1AEVZ_80MdI

The one variable had lower recall and accuracy score compared with our final three variable model. This means that we variables that we considered made our model more accurate as the accuracy score that we received for the second model was 15% higher than the first one.

***Conclusion:***

Our increase of variables considered made our model better as we had a better predictions of outcomes. The probability function was 75% accurate which is not the best we can do. Tests need to be done to verify which variables are optimizing the model and which are negatively affecting it as some variables could not be a good option to determine when is a pet getting adopted. Analyzing all permutations of variables with the result of adoption could help us achieve this. We confirmed that there are certain factors which are not random that determine if a pet is more likely to get adopted as we saw with the graph of age.

There are multiple improvements that can be done to this model. Some variables that are directly connected to the probability of a pet being adopted are health state and behavior. Health state is very important as not everyone can afford or handle an ill pet which will need special care and could be expensive to take care off, therefore the probability will most likely decrease. Behavior is also something that is directly connected as not everyone can handle an aggressive pet or would even want to have an aggressive pet. Most people that go into an animal shelter are looking for a family pet that can be around kids or other pets therefore this will greatly affect the probability of adoption for a pet. This model only considers dogs and cats as we did not have enough data for more pets to be taken into account for, but using data which could let us include birds, bunnies and other animals that an animal shelter accepts could help with the accuracy of it.

Acknowledgements:

* Tyler Le | Python Codding
* Sophia Lui, Alec Panattoni, Tyler Tran | Data Cleaning

References

1. Rossos, Katrina. “The Challenges and Benefits of Adopting a Senior Dog.” *BeChewy*, 15 June 2021, https://be.chewy.com/the-challenges-and-benefits-of-adopting-a-senior-dog-is-it-right-for-you/.
2. “What Kinds of Pets Get Adopted?” *Priceonomics*, https://priceonomics.com/what-kinds-of-pets-get-adopted/.
3. Powell, Lauren, et al. “Characterizing Unsuccessful Animal Adoptions: Age and Breed Predict the Likelihood of Return, Reasons for Return and Post-Return Outcomes.” *Nature News*, Nature Publishing Group, 13 Apr. 2021, https://www.nature.com/articles/s41598-021-87649-2.
4. Morgan, Liat, et al. “Human–Dog Relationships during the COVID-19 Pandemic: Booming Dog Adoption during Social Isolation.” *Nature News*, Nature Publishing Group, 24 Nov. 2020, https://www.nature.com/articles/s41599-020-00649-x.
5. County of Sonoma Department of Health Services. “Animal Shelter Intake and Outcome: Open Data: Sonoma County.” *Animal Shelter Intake and Outcome | Open Data | Sonoma County*, 3 Dec. 2021, https://data.sonomacounty.ca.gov/Government/Animal-Shelter-Intake-and-Outcome/924a-vesw.
6. “What Are the Different Dog Sizes? Dog Breeds Insider.” *Dogsized*, 11 Nov. 2021, https://dogsized.com/what-are-the-different-sizes-of-dogs/.
7. *Logistic Regression*, http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html.
8. “Installer Help Options.” *TIBCO Product Documentation*, https://docs.tibco.com/pub/sfire-dsc/6.6.0/doc/html/TIB\_sfire-dsc\_6.6.0\_install-admin/GUID-ECF5914C-BC9F-4DE4-9D65-FB138986C8F0.html.
9. “Logistic Regression.” *Logistic Regression - ML Glossary Documentation*, https://ml-cheatsheet.readthedocs.io/en/latest/logistic\_regression.html.
10. Kumawat, Dinesh. “Introduction to Logistic Regression - Sigmoid Function, Code Explanation.” *Analytics Steps*, https://www.analyticssteps.com/blogs/introduction-logistic-regression-sigmoid-function-code-explanation.