Autumn Salsberry
Capstone Report - August 8, 2022 - BrainStation

Predicting Adoptability of Shelter Dogs

Given the number of animal lovers around the world and the increase in pet adoption since the pandemic, there is a good chance either you or someone you know has been to an animal shelter. Though not everyone chooses to rescue their furry friends from shelters, even pets that are purchased can end up in the pound for various reasons. A lessor known fact about rescues is that they often are underfunded and over filled leading to the shelter needing to make the tough decision to euthanize dogs that are not likely to be adopted. Though this is tragic, it is the only realistic option for shelter staff that are stuck between a rock and a hard spot. They are in this line of work because they love animals, but they cannot reasonably care for all of them.

For this reason I had the idea of building a model that can accurately predict if a dog is likely to be adopted or not. Furthermore, this model would free up time and human resources by using an image of the rescued dog to predict the likelihood of it being adopted. The success of this model would mean shelters caould allocate their precious funds with a deeper understanding of which dogs had the most (or least) to gain. In the best case scenario, a dog that is unlikely to be adopted at one shelter likely could be adopted at another and thus the pup can be transferred to that facility freeing up a kennel for a dog more likely to be adopted in the first facility.

There have been amazing advancements in image classification over the past decade or so due to the increased knowledge of how our brains work. The many scholarly studies already put into image classification can be utilized to classify novel images and objects the models may have never seen. For this reason it is possible to train one of these models (that dozens of people have spent hundreds of hours fine tuning to optimize results for their projects) on my own image classification project. This process is called transfer learning. The proper classification of a dog image can be used to then predict if a dog is adoptable or not. This portion of the model will be built on real world data from a shelter interested in utilizing this model for their own management

purposes. Further research would need to be conducted to determine if such a model could be generalized to other shelters that did not provide data in the model building process.

With this in mind I sourced a large dataset (77,000 rows) from the Austin, TX animal shelter and an image data set (20,580 images) from Stanford University. The images were of 120 different breeds in varying natural setting, with breeds having varying coat colors and some images also containing humans. The Austin, TX dataset contained key details about how/when an animal came to the shelter, description of the pet (sex, age, breed, color, fixed, etc.), and key outcome details such as type (adopted, transferred, returned to owner) and date of outcome. This shelter data was collected for government census purposes and needed a lot of wrangling before any models could be built on it. Before cleaning the data I reduced it to only dogs since the image data I sources is only representative of dogs. I then analyzed each row individually and compared it to similar rows to eliminate redundancies. For example, there were intake dates in two different calendar formats. I changed most columns to binary format by analyzing trends and considering proportions. For example, there were about 10 times more strays than any other intake sub-category, so this column could be simplified to if a dog was rescued as a stray or not. The other kind of cleaning I did was OneHotEncoding the breed and coat color columns because they are so valuable to predicting if a pet is adoptable or not. Ideally, the image portion of the model would be identifying the breed and coat color. Then the model would use this information as input for predicting if a dog is likely to be adopted or not. Thus, it is important for the tabular data to have a good representation of this information. Of course I did also delete several irrelevant columns. As previously mentioned, this data is meant for the census, so there are several characteristics of a rescue that are useful for the shelter, but not for this model – for example the location of rescue. As for the images, I simply needed to split them into sets for training and testing the model, resize them so they all are the same number of pixels and read in the image pixel intensities to an array and create a matrix of

image arrays. This matrix of image arrays was then plugged into neural networks mostly built on the transfer learning models previously discussed.

Since transfer learning gives so much insight compared to the amount of effort required to build the model, the first half of the model (predicting the breed) was more simple than building the second half (predicting if the dog is likely to be adopted or not). The image portion of the model was 14 times better at predicting the breed of a dog as compared to chance. In order to optimize the second half of the overall model, I trained and fitted a wide variety of models including logistic regression, random forest classifier, neural networks, ada and xgboost. Of these models, logistic regression and random forest seemed to perform the best with a realistic amount of effort. They yielded a 19% increase on predictability compared with the base rate of dogs that were adopted divided by the total number of dogs.

Though these results are impressive, I don't think it would be possible to use both halves of the model as originally planned to predict adoptability of a dog given its image. The second half of the model could be used **as is** to help this shelter in Austin, TX make better informed decisions of their rescues dogs and resource allotment. With this information it is not possible to predict how much money and human resources the model could save, but having a 20% improvement on knowing if a dog will likely be adopted or not gives the management team an opportunity to build a heuristic based on their morals and needs.

As always there is room for improvement. With more research, fine tuning and even a larger dataset, it is possible to improve the image model to better predict the breed and potentially even recognize the color of the dog. These classifications can then be used as inputs for the model built on the shelter data and could finally predict if the dog is likely to be adopted or not. Given that this is a philanthropic project with important social benefits, I will continue to work on it as I learn and grow as a data scientist. I will publish these findings on Kaggle and GitHub and will likely reach out to the Austin, TX shelter and see if they could practically benefit from my findings.