**Capstone Project 1: Milestone Report**

For the first capstone project, I have been working with data from the AAC, which compiled shelter data on animals that were admitted/ released. This data included the specific breeds as well as the outcome, such as adopted or euthanized. This data was of interest as boosting adoptions is a goal of any animal shelter. The goal of my analysis was to see if there was a correlation between animal age and chance of being adopted. I also wanted to see if there were any breeds that resulted in higher adoptions. This is useful information to the client, AAC, as they can target specific breeds/ ages on their websites/ marketing in order to boost adoptions.

The dataset from AAC was already fairly clean with very clear column headers, but I did perform some wrangling steps to get the specific columns I needed. One of the first steps I took was creating datetime objects from the intake and outtake dates so that I could calculate the shelter length of each animal. I performed some exploration of the dataset and saw that there were numerous outcome types that did not seem sensical for AAC’s marketing purposes. For example, there was Return to Owner and Transfer as outcome options. These are not true cases that I wanted to review as they are not relevant to the adoption or euthanization outcomes, so I isolated only the outcomes that resulted in adoptions or euthanizations.

I also wanted to calculate the age of each animal upon outcome, so I converted the DOB column into a datetime object and subtracted from the outcome datetime. I also wanted to block the shelter length and age of outcome into meaningful buckets, so I used the time delta to convert the shelter length into weeks and age of outcome into years. This allowed those columns to be split into more sizable chunks that could be used for comparison. Once the age of outcome and shelter length columns were cleaned and converted, I combed through them and removed any outliers (there were 5 entries that had -1 as age of outcome so were removed). I also checked for any null values in the dataset and had these removed (there were 10 null values for outcome type, so just removed due to the low number).

Once the data was cleaned and ready for analysis, I began performing the EDA, which allowed me to see interesting trends in the dataset. Initially looking at the data, I noticed that adoptions outweighed cases of euthanization, but I also noticed that there seemed to be a trend with younger animals being adopted and euthanized, which was unusual to me. I expected older animals to lead the cases of euthanization, but that was not the case. At this point in the EDA, I wanted to look at breed specific trends so see if there were any patterns of use, especially with the cases of euthanization since they differed from my expectations.

Looking at the top breeds that were euthanized, I saw that Pit Bull Mixes ages 1-2 lead the charge for dogs; however, Domestic Shorthair Mix for cats more than tripled the dog amounts. The DS mix for cats led the top three spots for euthanizations when grouped by breed and age of outcome. I found that very strange, especially since the ages ranged from 0-2. I did not expect cats to lead the euthanization stats, so my next step was to review the top breeds for adoptions. I saw that the top spot was also the Domestic Shorthair Mix aged 0-1 with over 7500 cases. The second top breed was Labrador Retriever Mix aged 0-1 with over 1200 cases. I found that to be a shocking difference between the first and second top adopted breeds and pondered why there would be such a large discrepancy. It also seemed weird to me that cats accounted for the top number of adoptions because I find dogs to be more popular. As a cat owner myself, I usually feel in minority when discussing pets with friends or colleagues.

Thinking through the data, I came to the realization that many cat owners do not specify their cats’ breeds as most people do not know the distinct breeds of cats. People are familiar with the different dog breeds, so it’s very common for people to throw breeds around in conversation, but the same is not true for cats. I often refer to my cats as Domestic Shorthair because it is simply easier to say that rather than list out the actual breeds. After pondering that, I felt that cat data was actually throwing my numbers off and decided to solely focus on dogs for this project.

Once the cats were removed from the dataset, I looked at the top ten adoptions and euthanizations grouped by breed and age of outcome. I plotted histograms of this data to have a clearer picture of the spread. Interestingly, I saw that the top breed euthanized was the Pit Bull Mix ages 0-3, which seemed bizarre as the ages were so young. Looking at the adoptions, I saw that the Labrador Retriever Mix aged 0-1 accounted for the top adoptions by a landslide. The next breeds were Chihuahua Shorthair Mix (aged 0-1) and Pit Bull Mix (aged 1-2). The histogram was able to clearly show labrador retrievers accounted for about double the second highest breed adopted and then trailed off from there. The EDA showed interestly trends for the data that I wanted to explore further in the inferential statistics section.

The first thing I wanted to check was the correlation between age of outcome and adoption as I expected animals with higher ages led to longer shelter times; however, there was not much of a covariance between those two variables. From there, I calculated the shelter mean and STD so that I could perform a binomial distribution against the rate of adoptions against euthanizations. I used the np.randon.binomial function to achieve this, and plotted as a histogram. I saw that on average, there was a 93% chance of an animal being adopted and the sample data spanned from 85 - 97.5%

I also was able to compute the ECDF of this dataset and could see the majority of data was between the 87.5 and 97.5% mark. I then graphed the PMF of this sample distribution to create an alternative way of viewing the data. From there I wanted to perform poisson distributions of the shelter adoptions and calculate the mean as well as the percent of data less than or equal to 3 years. I saw that the mean shelter length was 3 and 64% of adoptions occurred for shelter lengths 3 and below. I then performed a poisson distribution of the age of adoptions and saw that the mean age was 2 and 73% of adoptions occurred for animals 2 and younger.

From my EDA, I learned that Pit Bull mixes are popular breeds for adoptions as well as euthanizations, I wanted to run a simulation on this. I created a new dataframe of Pit Bull mixes that were euthanized as well as adopted and then calculated the probability of pit bulls being adopted, which was 85%. From there, I conducted a binomial distribution on this dataset and saw that from the histogram and ECDF, 75% and 90% of pit bulls were adopted out of 100, so the majority of pit bulls are being adopted. I also performed a poisson distribution on the age of pit bulls euthanized and saw that the mean age of euthanized cases was 3 and 63% percent of pit bulls are euthanized between the ages of 3 and below. That is a high percentage for such a low age and made me wonder why so many young pit bulls are being euthanized. I performed the same poisson distribution for the age of adoptions and saw that the mean age was 2 and over 75% of pit bulls were adopted between the ages of 2 and below.

Since I saw that the labrador retriever mix was the highest adopted dog breed, I wanted to also perform a poisson distribution of this breed’s age of adoption for review. I saw that the mean age of adoption was 1 and 85% of labs were adopted between the ages of 2 and below. After that, I wanted to perform a binomial distribution of labs adopted out of the total population. I plotted this data as a histogram and ECDF and saw that spread was between 5 and 20%, which is the percentage of labs adopted out of the total population. I had expected this to be higher since Labs had such a large lead in the number of adoptions, but I am guessing this is due to the fact that labrador retriever mix is quite specific. Some people may just list labrador or retriever rather than actually labrador retriever mix.

I then wanted to review the amount of labrador retriever mixes euthanized verse adopted and performed a binomial distribution of the probability of labs being euthanized. From the distribution, I plotted the histogram and ECDF and saw that the data fell between 2 and 10%, which is very low and expected since this is the highest breed adopted for dogs. From the statistics performed, It seems to be in the AAC’s best interest to market labrador retrievers to boost adoptions and shy away from marketing pit bull mixes as there are an alarming amount euthanized between the ages of 0-3. The amount seems too high to account for illness so there must be another reason such as overpopulation or aggression. The data for the AAC does not provide a reason for euthanizing, but that would be helpful to take a deeper look into what led to these cases.

Throughout the machine learning section of my capstone, I experimented with different supervised learning techniques such as linear regression, ridge, lasso, and random forest ensemble. Based on the parameters of interest in my dataset, I wanted to focus on the age of adoption outcome and the shelter length to build a predictive model that could predict the shelter length based on the age of the dog. The first thing I did to get an idea of how these variables are related is plotted a scatter plot of shelter length versus age of outcome. I could see immediately that there are some outliers with shelter length and age, and the majority of the data is failing below 75 weeks of shelter length. After charting the data, I set-up my variables which were shelter length for the target and age of outcome for the feature. This is how I wanted the data to be set-up for modeling that way the model could predict the shelter length.

The first model I did was linear regression in Sklearn without any tuning. I chose this model as it is very easy to run and only requires a line or two of code. After modeling, I checked the score and saw it was fairly low, around 4.5%, which was concerning. I plotted a histogram of the predicted data and saw that the model was predicting the shelter length to be mainly 2-4 weeks, which was interesting because out true data is spread over 175 weeks. I then wanted to compare the predicted data against the original data set and plotted them together. This confirmed that the predicted data was representative of only 2-14 weeks of shelter length whereas our true data spans to 175 weeks. This showed me that the model is not very predictive of the dataset because the range is quite smaller. I also calculated the mean squared error and TSS. The MSE is calculating at 61, which is very high. This is showing that the model is underfitting the data, so wanted to explore other modeling choices.

From there, I ran an OLS model so that I could compare to the linear regression as well as review summary statistics. OLS is helpful because it provides good summary stats in just a few lines of code. From that data, I can also see the R square score is low, 4.5%; however, the probability is showing as 0, which is interesting because that does imply that the shelter length and age of outcome are statistically significant. I also graphed the predicted shelter length to compare to the linear regression data, and saw they were very similar. Next, I ran a ridge regression on the data and also scaled the data. I chose to run a ridge regression because in the Harvard lectures, there was a flow chart provided to help determine which tests to use based on data size and goals. Since I have a fairly large dataset and want to predict the shelter length, which is a quantity, this led me to run some models with Lasso and Ridge. This did improve the model slightly, but still underfitting. From there, I experimented with a lasso regression and also did not find that model fit the data, even with scaling. I decided to review my data again to see if I could remove any outliers that might be throwing the modeling off. I tried to remove shelter length greater than 50 and 75, but this really did not impact the data.

Next, I wanted to try a random forest ensemble thinking that it might be a better model for this data set as ensembles tend to fit data better as they are using an average. I ran the model with little to no tuning at first and got a mean squared error of 39, which was an improvement but still high. The R squared was still in the 4 range, which is very low. I played around with different tuning parameters of the random forest such as scaling, and altering the number of estimators. I still was not getting a significant increase in r squared. I decided to play around with my data again to remove ages that seemed outliers, but this still was not impacting. I also read articles on r squared to get a better understanding on what low r squared values mean for my models, and if this is something that I can improve with this data-set. I read that sometimes the data is just inherently unpredictable, especially involving people, so that can result in difficulties getting a high r squared value. After reading that, I thought that would make sense with this dataset because adoptions are unpredictable. Sometimes people want to adopt older animals, although that does not seem like the norm, I know many people who prefer adopting an older animal as they are already trained.

Next, I decided to pick out the models that were the most successful and fine tune them to see if that could drive the r square value up. My top performing models were ridge, linear regression, and random forest. After tuning the ridge regression, the best I could get was 4.6 %. Graphing the data and predicted line, you can see the model is just not fitting the data as expected. I did the same with the random forest, and the best I could get was 4.1 % with the predicted model not fitting the data as expected. Lastly, the best I could get after tuning the linear regression was 6.2, and also the model did not adequately fit the data. Although I was able to make improvements on the model, I was still not satisfied with the R squared that was calculated.

I also wanted to incorporate the breeds into my modeling to see how those would play out. I added the breeds as a feature to the original dataset I was using. I then used get\_dummies to convert the categorical breed column into binary numbers so that these could be modeled. I then ran a simple linear regression after reshaping the datasets, but also did not get an R squared value that I expected. My value was very negative, showing that the model is not fitting the data. Lastly, I wanted to try binning the ages into categories to see if that would drive the R squared value up. At first, I binned the age column into 0-1 year and 1 year and over. I used binning to create that new column and then used get dummies to transform that categorical column into binary numbers. After modeling, I still received a fairly low R squared score of 3.5%. I also graphed the predicted data to see what binning the ages did, and I can see that the spread of shelter length for dogs 0-1 is 0-50 weeks, which is a pretty large spread compared to the animals ages 1 and beyond, as those are only spreading to 100 weeks.

I also tried to play around with the binning, such as changing the age to 0-2 years. This did improve the model and R squared score. I was able to improve the R squared score from 4% to 4.5 with tuning. I also graphed the predicted data and original dataset. What is interesting is in the original data set, you can see the shelter length ranges from 0 -100 for animals 0-2 years, which is shocking because the common belief is that young animals should be adopted fairly quickly, but we are actually seeing differently in our dataset and thus the models themselves. It does not seem that the shelter length and age of outcome are as correlated as I had previously believed.

At this point, I think the models are underfitting the data, and I believe this may be because the relationship between the age and shelter length is complex and not as correlated as I had previously thought. I was able to improve the model through trial and error of different regression tests and hypertuning. The most successful model I was able to develop was using train test split, scaling and normalizing the data, and then performing a linear regression. This was the most successful at 6.2%. Although the model was not as predicative as I had initially hoped for when starting this section of my capstone project, I feel as though I learned a lot and was able to explore many different regression techniques. Also, since my model had not performed as expected, I did a lot of trial and error trying to fix and understand the behavior, which I do find helpful in retrospect.

Throughout the different stages of my capstone, I was able to learn and apply different skills ranging from EDA, to inferential statistics, story telling, as well as machine learning. I also was able to learn from some challenges that I had faced, specifically within the last section. I had not realized how important it was to evaluate how well different variables were correlated when I was performing the inferential statistics section. I had glossed over that and wanted to focus on different probabilities and parameters of interest. This in turn resulted in challenges faced in the machine learning section, as my shelter length and age of outcome are not as correlated as I was hoping for. Going into my second capstone project, I will definitely be on the lookout for that in my statistics section and focus on variables of interest that are also correlated.

Based on the testing and modeling I had done in this capstone, I think the AAC’s take-aways should be to not rely on ages of animals to predict their shelter length/ adoption rate as these were not very correlated as shown in the modeling. As far as advertisements go, I would not focus on specific ages to boost adoptions as the age that animals are being adopted is inconsistent. As far as breeds go, I would recommend shying away from advertisements with Pit Bulls. Oddly, I saw that this breed has high euthanization rates for puppies aged 0-3, so there seems to be some underlying relationship with this breed. As also seen during the inferential statistics section, I would recommend boosting advertisements of Labrador Retriever mixes as those have such a high adoption rate and low euthanization rate. Some follow-up questions and next steps for this capstone that I had were to explore why young Pit Bulls are being euthanized. The dataset did not have any further details on why a given animal was euthanized, but seems very odd that so many young puppies for that specific breed were euthanized. Also, as the modeling was not as predictive as I would have liked, I would like to run some modeling on the cats to see if their numbers were better at predicting shelter length from age. Even though their breeds did not seem accurate, their ages and shelter length should still be maintained. It would be interesting to see how the cat data modeled in comparison to the dog data.

All in all, I thought this capstone was very informative and have many take-aways to consider when picking a data set for my second capstone. I was able to apply what I had learned throughout the course material to a real-world problem, and look forward to seeing what the second capstone will bring.