Springboard Data Analytics Course

Pet Adoption Milestone Report for Capstone I

July 2019

John L. Parsons

Contents

1. Introduction

1.1 Problem Statement

The PetAdoption Dataset will have three business objectives. The first objective will focus on the cost for keeping shelters open to shelter pets. The cost between shelters varies, but is around 15 dollars per day to shelter a dog or cat and 200 dollars for the initial veterinarian bill (deworming, vaccinations, neutering and spaying) from the Foothills Animal Shelter for the United States. Figure 2 shows the distribution of animals for the five ordinal target variables of **AdoptionSpeed** in this study. The goal is to significantly reduce the 4,197 animals that stayed in the shelter for more than 100 days. Each day these 4,197 of animals are living in the shelters, the cost to maintain the shelters is around 62,955 per day. This objective is to find the champion model to predict which animals are the most and least likely to be adopted from the shelters. This information will be used in the Adoption Campaign make the necessary changes to minimize the animal's length of time in the shelter. The money saved in decreasing shelter time can be used for educating the public about the importance of spaying and neutering pets to reduce the number of unwanted animals, initiate a neutering and spaying campaign to help those that cannot afford this procedure and create a positive relationship with the community to help reduce the number of unwanted pets.

The second objective is to build a Survival Analysis Recommendation system. The goal is to develop a system that can rank shelter animals which are the least likely to be adopted from the shelters or high survival rates (more than 100 days in a shelter) and target these animals first for adoption to potential adoptees. The models can predict which animals are most likely to stay in the shelter for more than 100 days. The goal is to increase the chances for these animals to have a permanent home that are often overlooked in the adoption process and reducing the long-term shelter costs.

The third goal is finding the names of the pet that will increase an animal's chance of being adopted from the shelters. **Name** is not only a noun that can define a pet, but it may be a one-word advertisement or a "Framing Affect" that can draw a potential adoptee to the shelter to find out more about a particular pet. The text analysis will determine which names are the best predictors for **AdoptionSpeed**. Text analysis can also flag **Names** that should not be included in the pet description because they may reduce an animal's chance for being adopted.

1.2 Client

PetFinder was founded by Betsy and Jared Saul in 1996 in Pittstown, NJ. This became the largest online Pet Adoption company that has listed over 350,000 adoptees from 14,000 shelters across the world. PetFinder was purchased by the Nestle Purina Pet Care Company in June of 2013 and was the first major acquisition of a digital company. These business goals are targeted to decrease shelter time for pets, which will benefit the board of directors, investors, 18,000 employees and 400 scientists who work for the Nestle Petcare company to continue achieving their goals and make this a successful campaign for people and pets.

1.3 Dataset Summary

The Pet Adoption Dataset will be used for this project and came from the Kaggle Competition website (https://www.kaggle.com/c/petfinder-adoption-prediction). The goal of the competition was to predict the length of time animals stayed in the shelter and this was labeled as **AdoptionSpeed**. The **AdoptionSpeed** of the animals in the shelters was then predicted from a group of two files. These files were the training and pet images file. There were additional Excel CSV files for the description of the coat color, breed and state information and the test file for making the final prediction. The goal for the competition is to develop the best machine learning algorithm that can predict **AdoptionSpeed** of animals being adopted out of the shelter. The winner of the competition was [ods.ai] bestpetting with an overall score of 0.46613.

The training dataset will only be used for this project to meet the business objectives. The Pet Adoption training dataset contains a total of 24 attributes and 14,993 rows of data. The list of all 24 attributes can be seen in Figure 1. The PetAdoption dataset had three types of data classifications listed in Python. The majority of these were Int64 for 19 variables, four were listed as Objects (**Name, Description, PetID** and **RescueID**) and only one was listed as a Float64 (**PhotoAmt**). A complete description can be seen in Figure 3.

The **AdoptionSpeed** attribute is an ordinal variable that shows the length of time an animal stays in the shelter and is the target variable. The **AdoptionSpeed** variable has a total of five levels based on the age of the animal entering the shelter and the description of the levels can be seen in Figure 2. There were no animals that stayed in the shelter between 90 and 100 days.

Figure 1: List of all 24 Attributes found in the Pet Adoption data set.

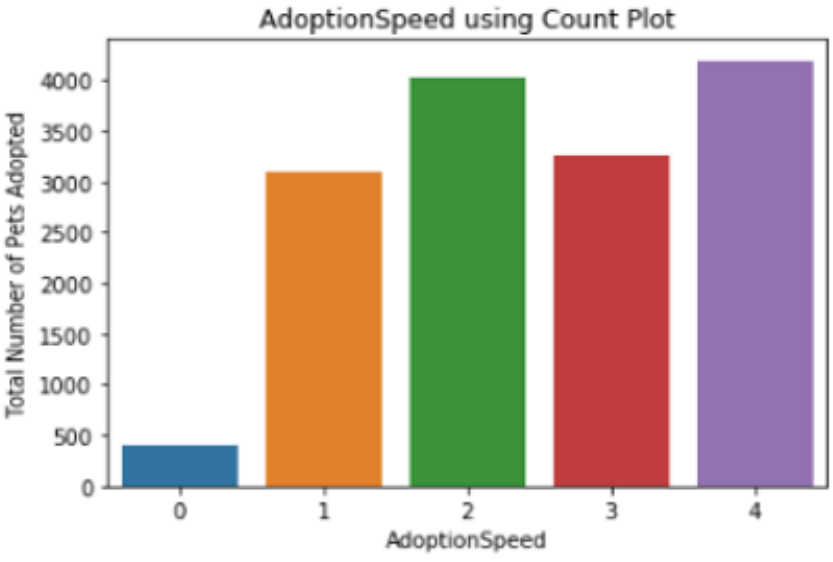
Figure 1: Pet Adoption Attributes 1 to 14.

	Type	Name	Age	Breed1	Breed2	Gender	Color1	Color2	Color3	MaturitySize	...	Health	Quantity
1	2	No Name Yet	1	265	0	1	1	2	0	2	...	1	1
2	1	Brisco	1	307	0	1	2	7	0	2	...	1	1

Figure 1 (continued): Pet Adoption Attributes 15 to 24.

Quantity	Fee	State	RescuerID	VideoAmt	Description	PetID	PhotoAmt	AdoptionSpeed
1	0	41401	3082c7125d8fb66f7dd4bff4192c8b14	0	I just found it alone yesterday near my apartm...	6296e909a	2.0	0
1	0	41326	fa90fa5b1ee11c86938398b60abc32cb	0	Their pregnant mother was dumped by her irresp...	3422e4906	7.0	3

Figure 2: Count and description of the AdoptionSpeed Variable.



Pet Adoption Category	Description of the Level
Level 0	Pet Adopted the same day it came into the shelter
Level 1	Pet Adopted between 1 and 7 days in the shelter
Level 2	Pet Adotped between 8 and 30 days in the shelter
Level 3	Pet Adopted between 31 and 90 days in the shelter
Level 4	Pet Adopted after more than 100 days in the shelter

Figure 3: Pet Adoption Variable Description.

Description of the Pet Adoption Variables
Data Fields PetID - Unique hash ID of pet profile
AdoptionSpeed - Categorical speed of adoption.
Type - Type of animal (1 = Dog, 2 = Cat)
Name - Name of pet (Empty if not named)
Age - Age of pet when listed, in months
Breed1 - Primary breed of pet is a Nomimal Variable with 176 Levels
Breed2 - Secondary breed of pet, if pet is of mixed breed is a Nominal Variable with 135 Levels
Gender - Gender of pet (1 = Male, 2 = Female, 3 = Mixed, if profile represents group of pets)
Color1 - Color 1 of pet is an Nomimal Variable with 7 levels
Color2 - Color 2 of pet is an Nominal Variable with 7 levels
Color3 - Color 3 of pet is an Nominal Variable with 6 Levels
MaturitySize - Size at maturity (1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large, 0 = Not Specified)
FurLength - Fur length (1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified)
Vaccinated - Pet has been vaccinated (1 = Yes, 2 = No, 3 = Not Sure)
Dewormed - Pet has been dewormed (1 = Yes, 2 = No, 3 = Not Sure)
Sterilized - Pet has been spayed / neutered (1 = Yes, 2 = No, 3 = Not Sure)
Health - Health Condition (1 = Healthy, 2 = Minor Injury, 3 = Serious Injury, 0 = Not Specified)
Quantity - Number of pets represented in profile
Fee - Adoption fee (0 = Free)
State - State location in Malaysia is a Nominal Variable with 14 Levels

2. Data Wrangling

2.1 Missing Data Points

The training data set did not have any missing values for the numeric variables in the dataset. There were two non-numeric variables that contained missing values and this was the **Name** and the **Description** of the pet. The **Name** variable had a total of 1,257 missing values and the **Description** only had 12 missing values. The missing rows will be removed for the text analysis section of this study.

2.2 Outliers for Selected Input Variables

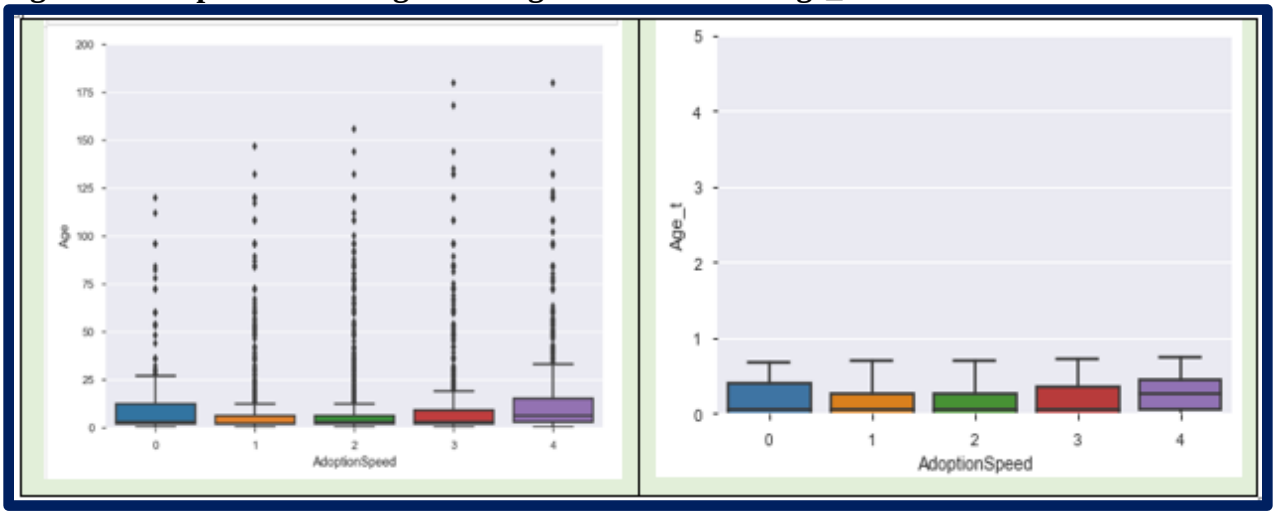
Bivariate boxplots were set up with the **Age** and **Breed1** variables and this shows several outliers for these input attributes. A total of 9,061 out of 14,993 pets were adopted that were listed between 0 and 5 days old when entering the shelter. The first seven-day results can be seen in the Cross Tabulations output in Figure 4. The **Age** attribute will be Log10 transformed to minimize the number of outliers and Figure 5 shows the bivariate boxplot for the **Age** and Log10

transformed **Age_t variables**. The Log10 values that are classified as "**inf**" will be given a value of 1 to complete the Machine Learning Models.

Figure 4: Cross Tabulations for Age (0 to 7) and AdoptionSpeed

Age / AdoptionSpeed	0	1	2	3	4	5	6	7
0	10	54	115	44	24	7	20	7
1	50	643	865	408	211	100	80	43
2	44	754	1120	586	265	157	117	59
3	39	511	783	458	260	135	115	62
4	36	342	620	470	349	196	226	110
All	179	2304	3503	1966	1109	595	558	281

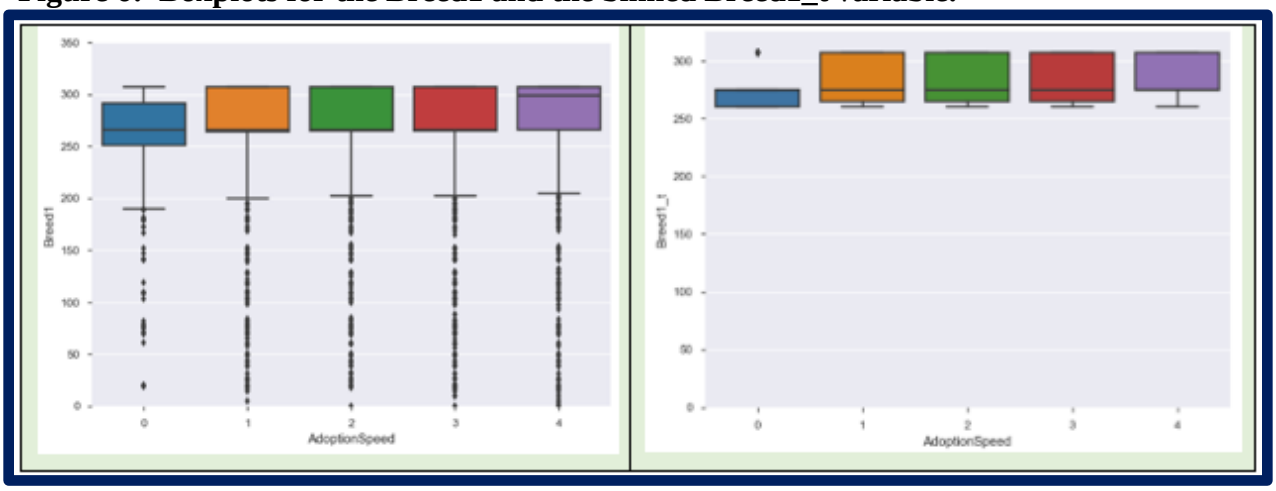
Figure 5: Boxplots for the Age and Log10 Transformed Age_t variable



The **Breed1** Attribute had a total of 176 unique factors for this variable and the bivariate boxplot shows several outliers. There were two breeds that dominated the number of pets in these agencies and this is 307 which accounted for 5,927 of animals and 266 which accounted for 3,634 of the animals. Breed 307 was a Tuxedo and 266 was a Domestic Short Hair animal. Breeds 266 and 307 remained as one category and then the remaining values were placed into two groups. All breeds that were less than or equal to 265 were placed in the bin named 265 and

all breeds greater than and equal to 267 (except for 307) were placed in another bin called 275. The results can be seen in the boxplots for the **Breed1** and binned **Breed1_t** variable in Figure 6.

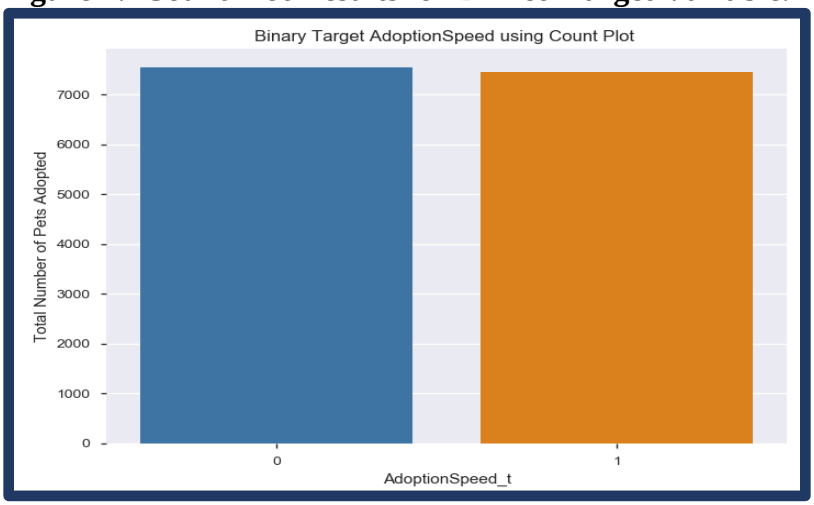
Figure 6: Boxplots for the Breed1 and the binned Breed1_t variable.



2.3 Handling Multifactorial Categorical Variables

The data wrangling step for the target variable will merge the five-level ordinal variable to a binary variable and will be used for the optimized model. The target variable will have levels 0, 1 and 2 binned into a 0-level and is for any animal that has been in a kennel for less than or equal to 30 days. The target variable with levels 3 and 4 will be binned into a 1 category for animals that has been in the shelter for more than 30 days. The binned Target variable results can be seen in Figure 7.

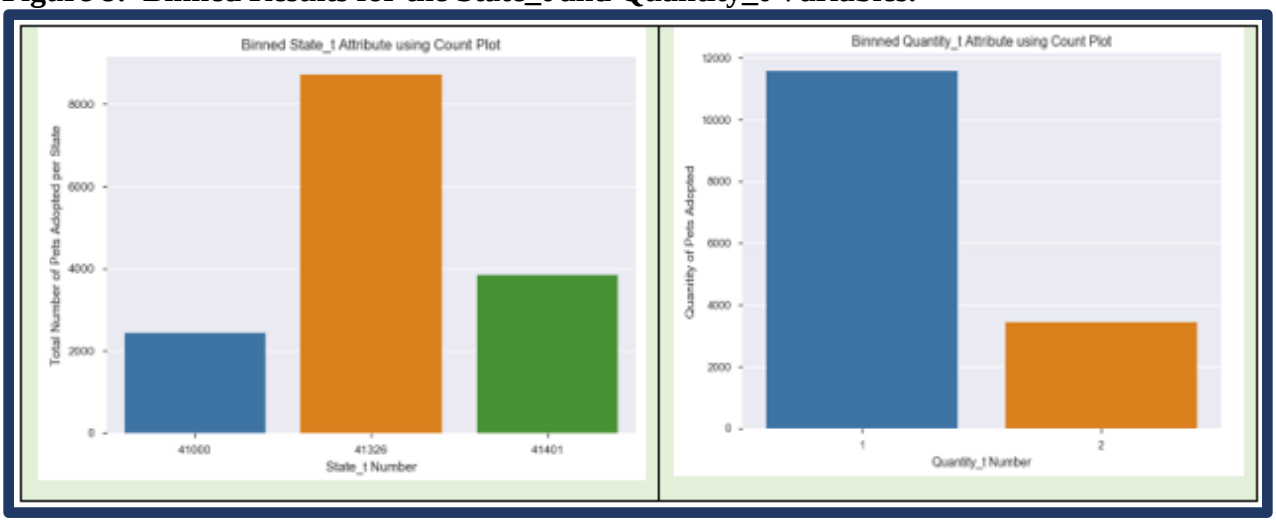
Figure 7: Count Plot Results for Binned Target Variable.



There are two input variables that had more than 7 levels for this study and are of interest for the optimized models. These two input variables are the **State** or location in Malaysia and

Quantity or number of pets represented in the profile for each kennel. The **State** variable had a total of 14 nominal levels and 41326 accounted for a total of 8,714 of the pets and 41401 accounted for 3,845 of the pets. Number 41326 is from Selangor and 41401 is from Kuala Lumpur in Malaysia. The remaining 12 levels were merged into one group called 41000 and the binned count for this variable can be seen in Figure 8. The **Quantity** ordinal variable had a total of 19 levels and range from 1 to 19 for this dataset. The greatest number of adoptions occurred when only one pet was listed at a time and this accounted for 11,565 of these pets and 1,422 for two pets listed at a time. These numbers continually decreased when two or more pets were included for the listing of these shelter pets. The **Quantity** variable will bin all animals that had 2 or more animals for a listing for group 2 and this can be seen in Figure 8 below.

Figure 8: Binned Results for the State_t and Quantity_t Variables.



3. Exploratory Data Analysis (EDA)

3.1 Questions of Interest

There was a total of four variables that were tested to determine if there were no significant differences between the attribute levels and the target variable and these are **Gender, Type, Sterilized and Vaccinated**. The standard Ordinary Least Squares was run for all four attributes and there were no dummy variables created so the Intercept is the mean for the first group in the dataset. The comparison of the means for Tukey was also run in Python and the results show if the null hypothesis can be rejected at a 0.05 p value. The statistical results from selected categories can be seen in Figures 9 below.

Figure 9: Select Statistical Results from the OLS and Tukey Models in Python.

Variable	OLS Results in Python					Tukey Results in Python					
	Group	P>|t|	F-Stat	R-squared		Group1	Group 2	meandiff	lower	upper	Reject @ 0.05
Gender	Male	0.000	30.5	0.004		1	2	0.1512	0.1020	0.2003	TRUE
Gender	Female	0.000	30.5	0.004		1	3	0.1675	0.0979	0.2372	TRUE
Gender	Mixed	0.000	30.5	0.004		2	3	0.0164	-0.0509	0.0836	FALSE
Type	Dog	0.000	125.8	0.008		1	2	-0.2156	-0.2533	-0.1779	TRUE
Type	Cat	0.000	125.8	0.008							
Sterilized	Yes	0.000	290.5	0.037		1	2	-0.5406	-0.5962	-0.4850	TRUE
Sterilized	No	0.000	290.5	0.037		1	3	-0.1860	-0.2661	-0.1060	TRUE
Sterilized	Not Sure	0.000	290.5	0.037		2	3	0.3545	0.2855	0.4236	TRUE
Vaccinated	Yes	0.000	122.9	0.016		1	2	-0.3081	-0.3561	-0.2600	TRUE
Vaccinated	No	0.000	122.9	0.016		1	3	-0.0410	-0.1137	0.0316	FALSE
Vaccinated	Not Sure	0.000	122.9	0.016		2	3	0.2670	0.1960	0.3381	TRUE

The results show a significant difference in the means between the male and female pets and the male and mixed lot pets. The null hypothesis can be rejected and in turn accept the alternative hypothesis for both groups means. The bar plot in Figure 10 shows more cats were adopted from the shelter for all five categories compared to the male pets and the mixed group came in a distant third for every category when compared to both levels. Figure 22 is a survivorship curve for the Cox Model and show cats had the lowest survivorship (length of time in the shelter) compared to the males from Day 1 to Day 250.

The **Type** and **Sterilized** variables were also significant for all levels and the null hypothesis could be rejected and accept the alternative Hypothesis. The graph in figure 10 shows that the cats were adopted the most for the first two AdoptionSpeed levels (0 to 7 Days) and then dogs were adopted the most for the remaining three levels for the target variable. The **Sterilized** category was interesting in this study because the unsterilized animals were the most adopted and sterilized animals coming in second place for being the most adopted with the AdoptionSpeed levels greater than or equal to 2 (Figure 11). Sterilization is an important factor in preventing unwanted pets and is an area that needs to be addressed in this study.

The **Vaccinated** attribute was the last attribute to test and the results show the null hypothesis could be rejected for pets that were vaccinated to those that were not vaccinated and also for the animals that were not vaccinated to those which did not have records for their vaccinations. The unvaccinated animals were adopted the most for the first four levels and then vaccinated animals were the most adopted for all animals that had been in the shelter for more

than 100 days. The number of pets who were vaccinated increased in the number of adoptions from the first to last level except for the third level, which had fewer adoptions than the previous level (Figure 11).

Figure 10: Bar Count plots for Gender and Type Variables

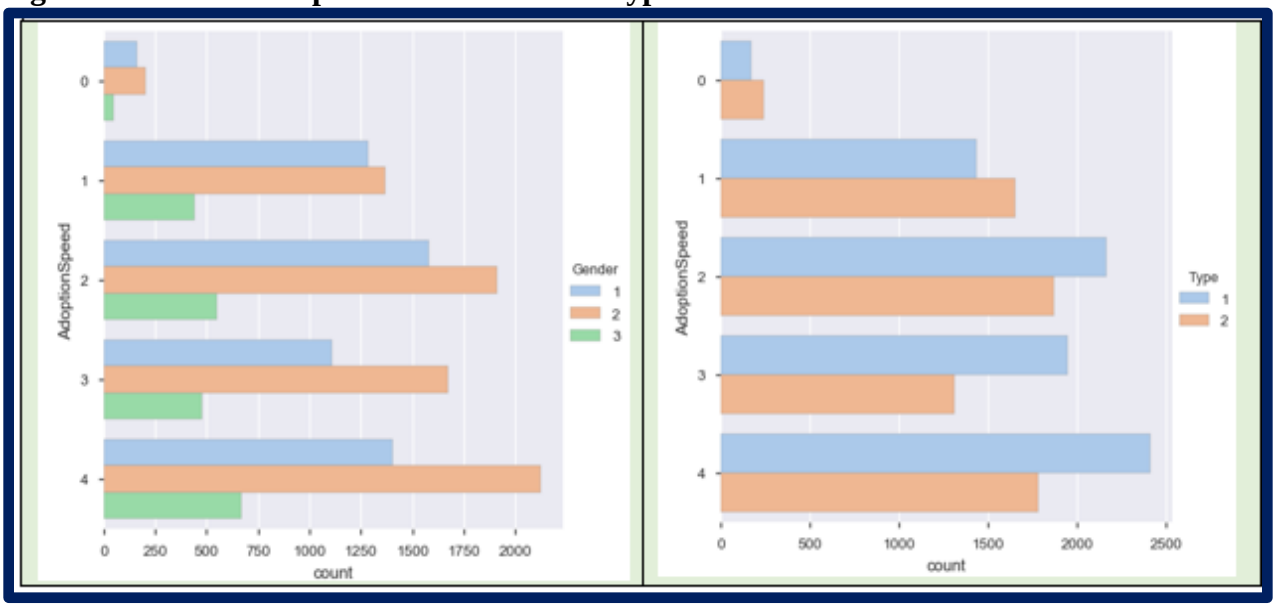
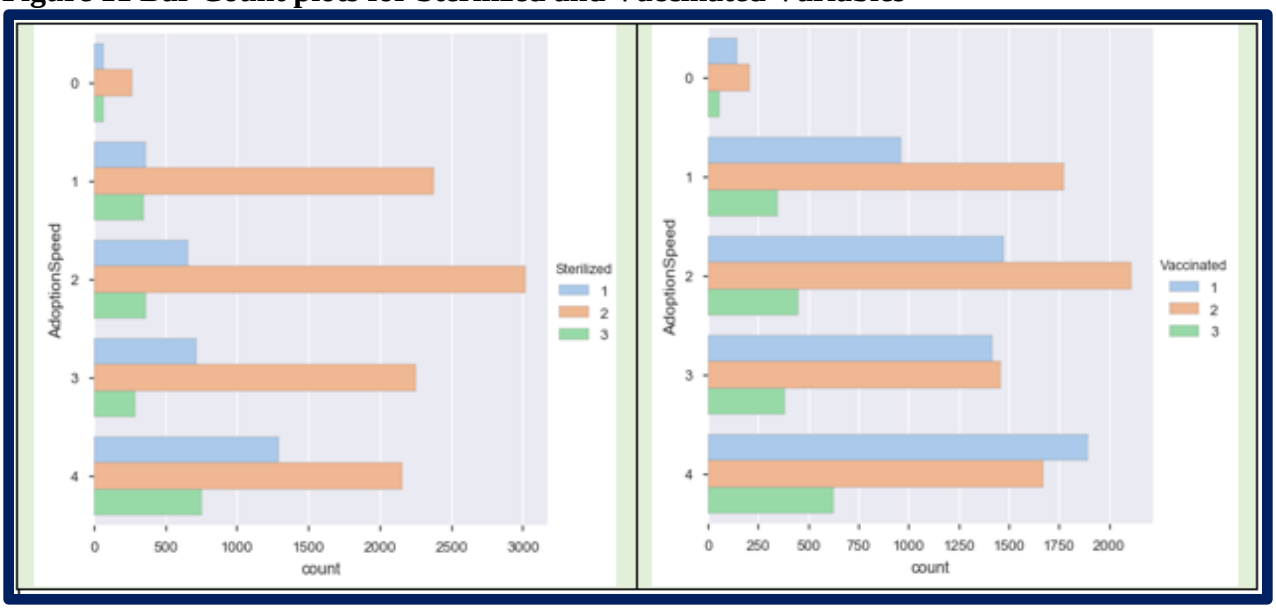


Figure 11 Bar Count plots for Sterilized and Vaccinated Variables

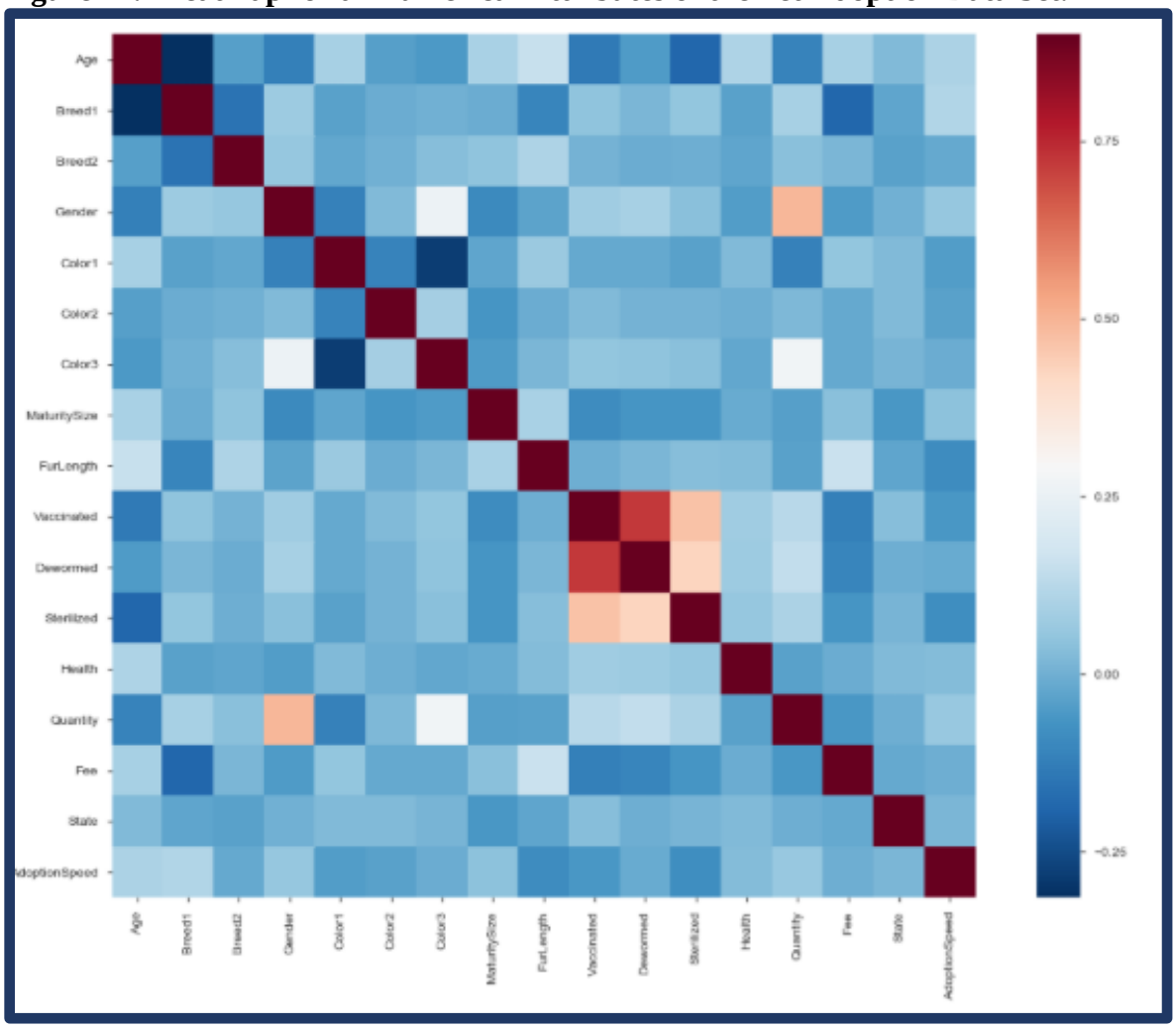


3.2 Correlation Heatmap

The Correlation Heatmap Matrix can be seen in Figure 12 below for the numerical attributes. **Breed1** and **Age** attribute has a slight negative correlation with a value -0.314.

Color1 and **Color 3** also had a slight negative correlation with value of -0.282. Three attributes had a more positive correlation and these are **Vaccinated, Dewormed and Sterilized**. The **Vaccinated** and **Sterilized** was 0.47 and the **Dewormed** and **Sterilized** was 0.436 for the Pearson Correlation Matrix. The two input variables which had the highest correlation are the **Vaccinated** and **Dewormed** variables and will be discussed next for the multicollinearity section.

Figure 12: Heatmap for all numerical Attributes of the Pet Adoption Data Set.



3.3 Multicollinearity

The one pair of input attributes that had the highest Pearson correlation values in the matrix was **Vaccinated** and **Dewormed**. These independent variables seem to be related to each other or have multicollinearity and the results can be seen in Figure 13. When animals are

brought in to the veterinarian for a checkup, this may be a common procedure done to the shelter pets. The final model will only select one of the two attributes to be used for the champion model.

Figure 13: Pearson Correlation Values

	Vaccinated	Dewormed
Vaccinated	1.000000	0.722596
Dewormed	0.722596	1.000000

3.4 Variable Importance

All 17 input variables will be used to create the baseline model for this study. When the baseline model has been developed, the best attributes will be selected using Pearson Correlations, p-values, Logworth values and Relative Importance values. A total of eight attributes have been selected to improve the model performance and these are; **Age, Breed1, FurLength, Quantity, Vaccinated, Gender, MaturitySize and Sterilized.** The Relative Importance Values from several models actually had **Color1**, **Color2**, **Fee and State** as an important variable and will be added later to see if this improves the model's performance. The goal is to find the best predictors for model performance that can be used to increase the success of the model and provide key insights into Pet **AdoptionSpeed** (Figure 14).

Figure 14: Correlations Plot Values and Logworth Values from JMP.

Correlations Values

	Abs Value
Breed1	0.122706
Age	0.092442
Sterilized	0.078757
FurLength	0.073397
Vaccinated	0.060014
Gender	0.058526
Quantity	0.046275
MaturitySize	0.039027
Color1	0.037945
Health	0.030723
Color2	0.029084
Breed2	0.016361
Dewormed	0.016070
Color3	0.007004
Fee	0.005864
State	0.004740

Logworth and RSquare values using JMP in SAS

Target Variable	Attribute	Count	PValue	LogWorth	FDR LogWorth	RSquare
AdoptionSpeed	Breed1	14993	5.06383E-40	39.2955212	38.06507231	0.011628
AdoptionSpeed	Age	14993	5.70687E-35	34.2436019	33.31418294	0.010102
AdoptionSpeed	Type	14993	4.34343E-29	28.3621675	27.73377862	0.008325
AdoptionSpeed	FurLength	14993	3.79565E-29	28.4207141	27.73377862	0.008342
AdoptionSpeed	Sterilized	14993	1.38603E-24	23.8582288	23.32674991	0.006963
AdoptionSpeed	Quantity	14993	1.23178E-14	13.9094661	13.45716839	0.00396
AdoptionSpeed	Vaccinated	14993	4.64665E-13	12.3328602	11.94750927	0.003486
AdoptionSpeed	Gender	14993	1.65539E-12	11.7810998	11.45374082	0.00332
AdoptionSpeed	MaturitySize	14993	2.23928E-08	7.64989214	7.373685729	0.002084
AdoptionSpeed	Color1	14993	6.18623E-08	7.20857409	6.978125167	0.001953
AdoptionSpeed	Color2	14993	2.29844E-06	5.63856603	5.449509797	0.001488
AdoptionSpeed	Health	14993	0.000320493	3.49418187	3.342914195	0.000863
AdoptionSpeed	Breed2	14993	0.022448763	1.64880758	1.532302015	0.000348
AdoptionSpeed	Dewormed	14993	0.10579015	0.97555477	0.911432288	0.000174
AdoptionSpeed	State	14993	0.108195712	0.96578995	0.911432288	0.000172
AdoptionSpeed	Color3	14993	0.390459723	0.40842376	0.382094819	4.92E-05
AdoptionSpeed	Fee	14993	0.62286522	0.20560592	0.205605919	1.61E-05

4. Preliminary Model Testing

The baseline model will use all 17 original numerical input variables to get a baseline result for the five-ordinal level of **AdoptionSpeed**. The standard Logistic Regression model will use C = 1 and a penalty of L1 for the model. The data was divided into a 65:35 split between the training and testing data. The results from this run can be seen in section 4.1 of this report when all 17 attributes were used for input variables. Section 4.2 will show the results with the same Logistic Regression when only the eight select input variables and a binary target variable was used for this model.

4.1 Baseline Logistic Regression with all Attributes

The results from the baseline model had low accuracy scores for the Logistic Regression. The Accuracy results from this model was at 33 percent and the Confusion Matrix can be seen in Figure 15. The baseline model did not classify any pets as adopted for the 0 level, but did a better job in classifying level 4. Level 4 had the best precision and recall with values of 0.36 and 0.67 (Figure 16). Level 2 had the second-best recall results with a score of 0.37 and third best precision estimates at 0.30. These values are low in predicting values to be positive that is actually positive or ones that are truly positive.

Figure 15: Confusion Matrix for the baseline model.

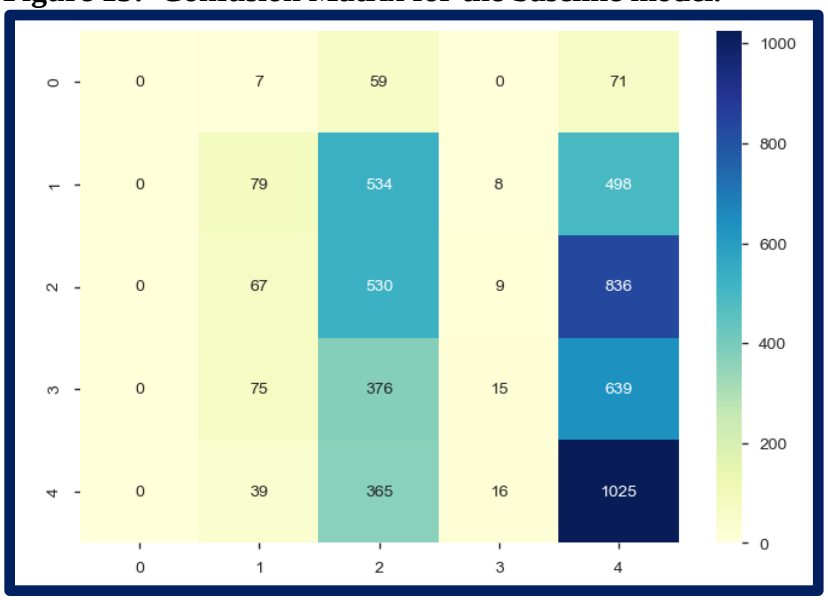


Figure 16: Classification Report for the baseline model

	precision	recall	f1-score	support
0	0.00	0.00	0.00	137
1	0.33	0.16	0.22	1119
2	0.30	0.37	0.33	1442
3	0.31	0.09	0.13	1105
4	0.36	0.65	0.47	1445
micro avg	0.33	0.33	0.33	5248
macro avg	0.26	0.25	0.23	5248
weighted avg	0.32	0.33	0.29	5248

4.2 Logistic Regression for the optimized model

A total of eight select input attributes were used for the first optimization step. These eight attributes are**; Breed1_t, Age, Type, FurLength, Sterilized, Quantity_t, Vaccinated and Gender**. The Target Variable **AdoptionSpeed** was binned into two levels and the **AdoptionSpeed**_t variable was used in the optimized model. The accuracy score was much higher than the baseline model with a value of 58%. The Confusion Matrix and Classification Report can be seen in Figures 17 and 18. The optimized model does a better job in classifying the True Positives (levels 0, 1 and 2) and True negatives (levels 3 and 4) compared to the baseline model. The Precision, Recall and F1 scores for this model ranged from 0.57 to 0.60 and this model is a significant improvement over the baseline model.

Figure 17: Classification Matrix for the Optimized Model.

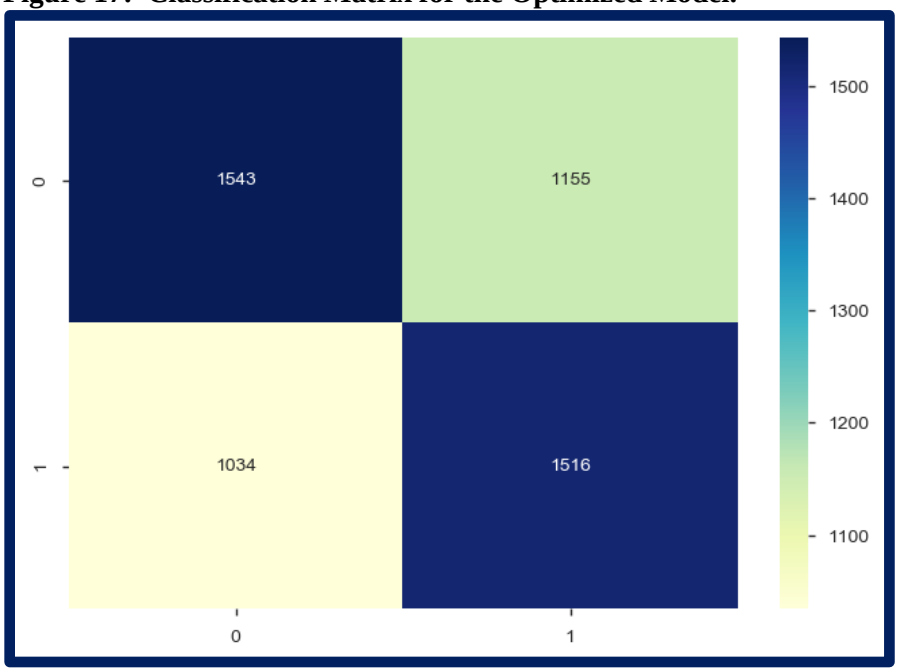


Figure 18: Classification Report for the optimized model.

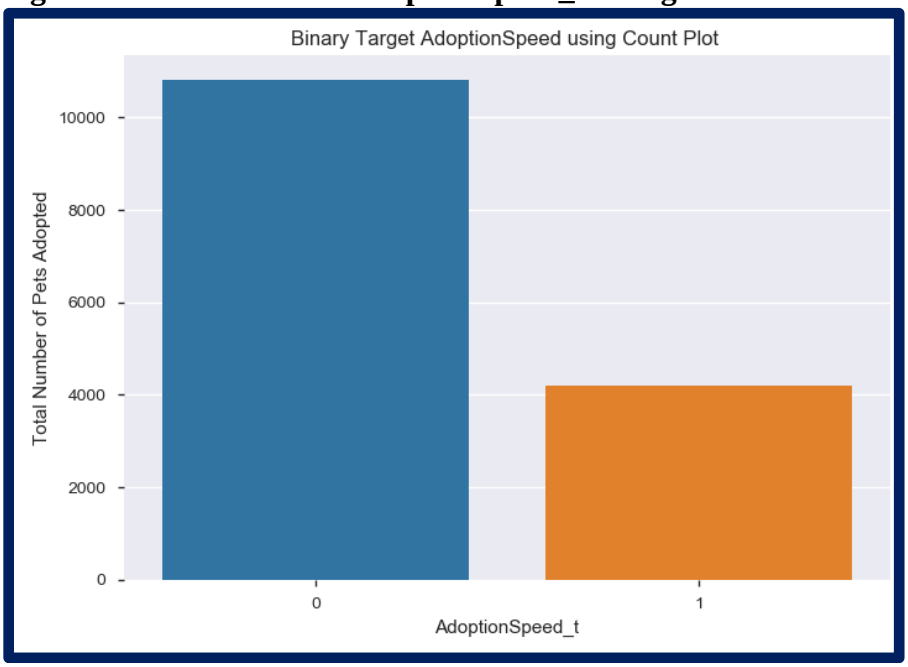
	precision	recall	f1-score	support
0	0.60	0.57	0.59	2698
1	0.57	0.59	0.58	2550
micro avg	0.58	0.58	0.58	5248
macro avg	0.58	0.58	0.58	5248
weighted avg	0.58	0.58	0.58	5248

4.3 Survivorship Model

The survival regression analysis will be used from the suite of Lifeline plots in Python. The goal for the survival regression models is to develop a model that can predict which pet is the least likely to be adopted that has been in the shelter for 100 or more days. The Cox model and Kaplan Meier Fitter Models will be used to develop models and create graphs for the Lifespans of select attributes. The top eight attributes will be used for these models and these are **Breed1_t, Age, Type, FurLength, Sterilized, Quantity_t, Vaccinated and Gender**. The Target Variable **AdoptionSpeed_t2** will be used for this analysis and levels 0 to 3 will be binned into the 0 level and level 4 will be binned into the 1 level (Figure 19). There was a total of 4,197 animals that remained in the shelter for more than 100 days and is the population of interest in the lifeline study. The goal is to

have a lower longevity value for survivorship in the model or adopted from the shelter. The Age attribute was selected for the duration column and the transformed AdoptionSpeed_t2 was selected for the event column.

Figure 19: Transformed AdoptionSpeed_t2 Target Variable Plot



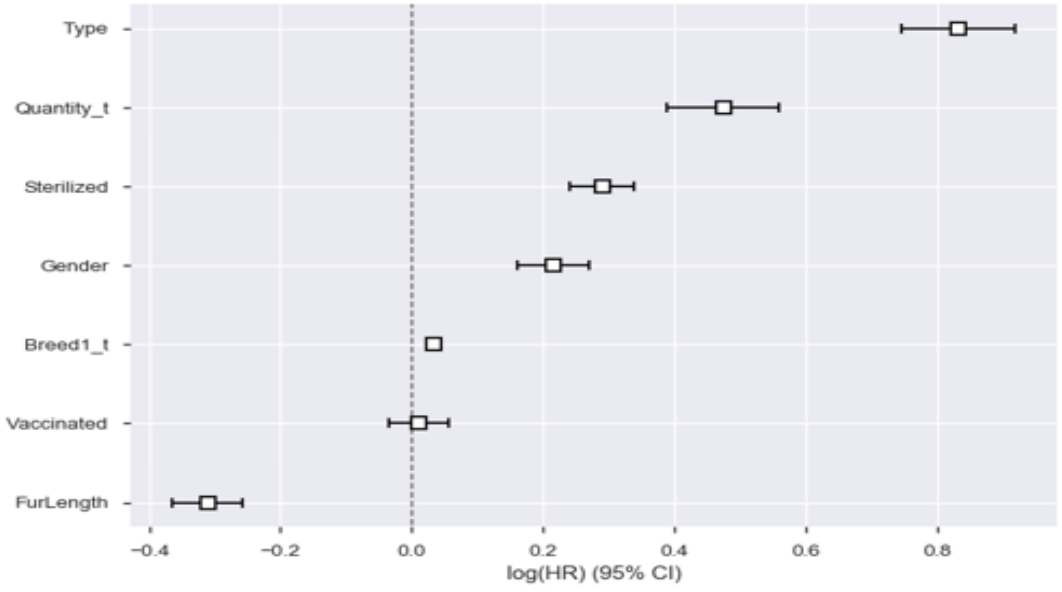
The summary output for the Cox coefficients and hazards functions shows seven of the eight attributes had a significant impact on AdoptionSpeed_t2 and Vaccinated was the only variable that did not have a significant affect with a p value of 0.64 (Figure 20). The FurLength attribute had the only negative impact for the baseline hazards in the shelter with a -0.31 value and all other variables had a positive impact on hazards function. The **Type** attribute had the largest positive hazard function with a value of 0.83. The goodness of fit value for the current model was at 72 % and did a good job in fitting the model to the select eight input variables (Figure 21).

Figure 20: Cox PHFitter Summary Statement for select inputs.

```
<lifelines.CoxPHFitter: fitted with 14993 observations, 10796 censored>
      duration col = 'Age'
         event col = 'AdoptionSpeed_t2'
number of subjects = 14993
  number of events = 4197
partial log-likelihood = -34271.87
  time fit was run = 2019-07-23 10:41:54 UTC

---
            coef exp(coef)  se(coef)      z      p  -log2(p)  lower 0.95  upper 0.95
Breed1_t    0.03      1.03      0.00  29.43 <0.005    629.89        0.03        0.04
Type        0.83      2.30      0.04  18.78 <0.005    258.86        0.75        0.92
FurLength  -0.31      0.73      0.03 -11.52 <0.005     99.66       -0.36       -0.26
Sterilized  0.29      1.34      0.03  11.39 <0.005     97.47        0.24        0.34
Quantity_t  0.47      1.61      0.04  10.97 <0.005     90.63        0.39        0.56
Vaccinated  0.01      1.01      0.02   0.47   0.64      0.64       -0.04        0.06
Gender      0.22      1.24      0.03   7.78 <0.005     47.00        0.16        0.27
---
Concordance = 0.72
Log-likelihood ratio test = 2089.36 on 7 df, -log2(p)=inf
```

Figure 21: Plotting the Coefficients for the Cox PHFitter Model.



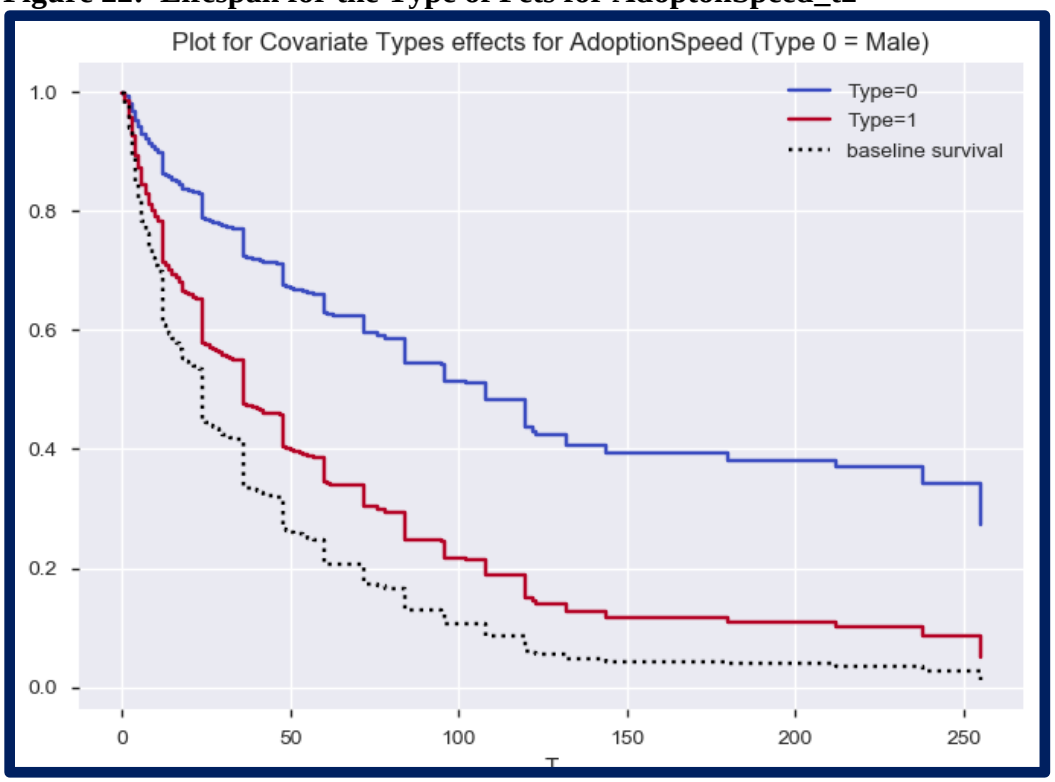

```
print(cph.score_)
```

```
0.720121369311449
```

The survival curve for the pet **Type** attribute was plotted against the AdoptionSpeed while holding all other parameters equal and the results can be seen in Figure 22. The curve shows that both types deviated from the baseline survival curve but the cat species had a lower

survival rate (adopted more frequently) than the dogs and begin to separate around the Age of 7 and then remained around 30% higher than dogs for survival from the age of 125.

Figure 22: Lifespan for the Type of Pets for AdoptonSpeed_t2



5. Text Analysis for Pet Names

The Text Analysis for the Pet Adoption Dataset will only look at the Names of the pets to calculate the term frequency, word count, polarity value and then use the Naïve Bayes and Random Forest to determine if Names can predict the AdoptionSpeed of the animal. The **Name** variable will have the standard text processing steps completed which are converting letters to lower case and removing special characters, whitespace, stopwords and most frequent words. The text will also use try stemming or Lemmatization of the words to improve model performance. The preprocessed text will then be used to calculate the length of the words, word count, polarity score, unigrams and bigrams.

A preliminary text analysis in python shows the top names which is more of an advertisement than a name. A list of the most Positive, Neutral and Negative ratings can be found in Figure 23. The polarity score was calculated and graphed and can be seen in Figure 24.

The majority of names were neutral in value with few names that scored high for the positive and negative sentiment in this data set. The names selected for the pets seemed to be more descriptive in nature and tell the potential adoptees the type of pet and physical description such as color, gender and circumstance (adoption or new home). There are also cultural differences in the names because the adoption shelters are from Malaysia and may impact the polarity score. There are few words like" Mimi" that are found in the top most common words. When Mimi is translated through Google translation this means noodle and is a noun for the Malay Language. This can be seen in Figure 25 which lists the top 20 common names for the data set.

Figure 23 Top Sentiment Polarity words for the Positive, Neutral and Negative Values.

Highest Positive Rating	Neutral Polarity Rating	Highest Negative Rating
faithful <3 the eight magnificent mongrel four awesome magnificent 7 bonbon a perfect companion 4 cat	maya jayden ti-ara don perry southern park kids	nameless tommy boy blind cat shy girl 4 kittens 1 blind & mummy

The article in Animal Farm Foundation (2013) discusses an interesting topic about the "Framing effect" for shelter pets. The Framing affect is how the information is presented such as a name and how it can influence decision making about that information for a specific type of pet. The goal in this study is to Frame the dogs in this article with names that make the potential adoptees feel good and positive about adopting a particular pet. They recommended using popular sounding names with minimal negative association. The majority of the names in this study have very neutral sentiment and the goal is to determine if the more positive names have a greater chance for being adopted when compared to the neutral or negative sentiment with the Naïve Bayes and Random Forest Models.

Figure 24 Polarity Score for Name.

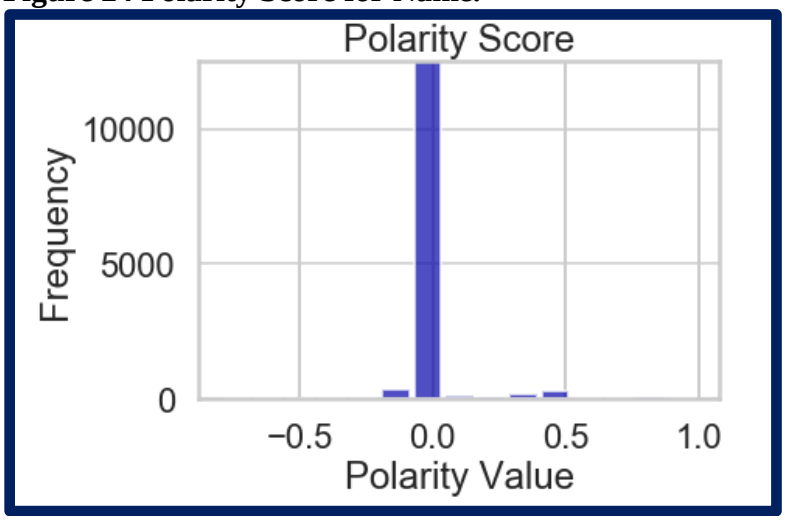
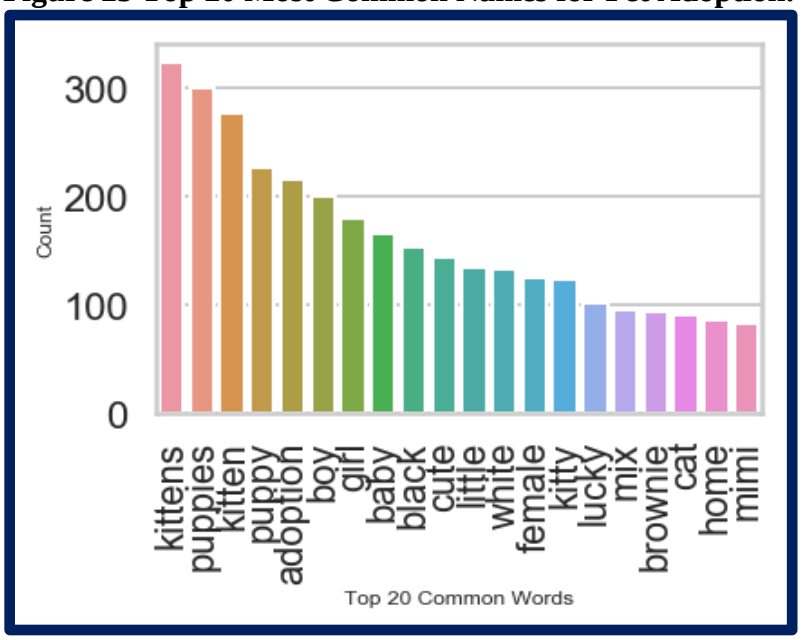


Figure 25 Top 20 Most Common Names for Pet Adoption.



A preliminary test was set up for the text analysis using the **Name** and **AdoptionSpeed** in the Naïve Bayes model. The standard preprocessing steps were performed (remove punctuation, stopwords and convert all letters to lower case) and the words were vectorized with the bow transformer approach in Python. The Naïve Bayes model tested the Name attribute with the five ordinal and the binomial level target variables. The accuracy results for the five-level target variable was 21% compared to 53% for the binomial level and the Classification Report can be seen in Figure 26 Below. The accuracy score for the text analysis in Naïve Bayes and the

Logistic Regression model for the Pet Adoption are similar in values and are currently weak predictors for the target variable.

Figure 26: Naive Bayes Model for testing Name Attribute with a Binomial Target Variable Classification Report.

```
[[1392  679]
 [1263  784]]


              precision    recall  f1-score   support

           0       0.52      0.67      0.59      2071
           1       0.54      0.38      0.45      2047

   micro avg       0.53      0.53      0.53      4118
   macro avg       0.53      0.53      0.52      4118
weighted avg       0.53      0.53      0.52      4118
```

6. Conclusion

The Pet Adoption Milestone Report for Springboard has listed the problem statement for pet adoption and three problems to address for this business objective. The three business goals are to reduce the cost in maintaining shelter pets by creating a champion machine learning model that can predict which animals are the most likely to be adopted and survivorship models to predict which animals that are least likely to leave the shelter. These animals will be targeted in the adoption process to market these individuals first to potential adoptees in attempt to lower long-term shelter costs for hard to adopt pets. The third business objective to do a text analysis for the name of the pets and determine which names have a higher adoption rate and can be used as a Framing Effect to be more successful in the adoption campaign.

The next set of deliverables will be the completion of the Capstone I project which will contain the champion model to predict AdoptionSpeed, the best survivorship model to predict which animals are most likely to stay in the shelters and the most efficient names to be used for pet adoption.

7. References

Animal Farm Foundation. (ND). Naming Shelter Dogs: The Framing Effect. Retrieved from: https://animalfarmfoundation.blog/2013/04/08/dog-names-framing/

Davidson-Pilon, C. (ND). Lifelines. Retrieved from: https://lifelines.readthedocs.io/en/latest/

Foothills Animal Shelter. (ND). Services; Lost and Found Fees. Retrieved from: https://foothillsanimalshelter.org/services/lost-and-found/fees/

Hastie et al. (2017). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Series in Statistics, 2nd Edition.

Purina. (2013). Nestle Purina completes acquisition of Petfinder. Retrieved from: http://newscenter.purina.com/2013-07-15-Nestle-Purina-completes-acquisition-of-Petfinder