Springboard Data Analytics Course

Pet Adoption Milestone Report for Capstone I

July 2019

John L. Parsons

Contents

[**1. Introduction** 3](#_Toc14763821)

[**1.1 Problem Statement** 3](#_Toc14763822)

[**1.2 Client** 4](#_Toc14763823)

[**1.3 Dataset Summary** 4](#_Toc14763824)

[**2. Data Wrangling** 6](#_Toc14763825)

[**2.1 Missing Data Points** 6](#_Toc14763826)

[**2.2 Outliers for Selected Input Variables** 6](#_Toc14763827)

[**2.3 Handling Multifactorial Categorical Variables** 8](#_Toc14763828)

[**3. Exploratory Data Analysis (EDA)** 9](#_Toc14763829)

[**3.1 Questions of Interest (4 Figures)** 9](#_Toc14763830)

[**3.2 Correlation Heatmap** 12](#_Toc14763831)

[**3.3 Multicollinearity** 13](#_Toc14763832)

[**3.4 Variable Importance** 14](#_Toc14763833)

[**4. Preliminary Model Testing** 15](#_Toc14763834)

[**4.1 Baseline Logistic Regression with all Attributes** 15](#_Toc14763835)

[**4.2 Logistic Regression for the optimized model** 16](#_Toc14763836)

[**4.3 Survivorship Model** 17](#_Toc14763837)

[**5.0 Text Analysis for Pet Names** 20](#_Toc14763838)

[**6.0 Conclusion** 23](#_Toc14763839)

[**7.0 References** 24](#_Toc14763840)

# **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) in the Foothills Animal Shelter for the United States. Figure 3 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 and have positive long term affects for the community.

The second objective is to build a Lifeline survival analysis system. The goal is to develop a system that can rank those 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.

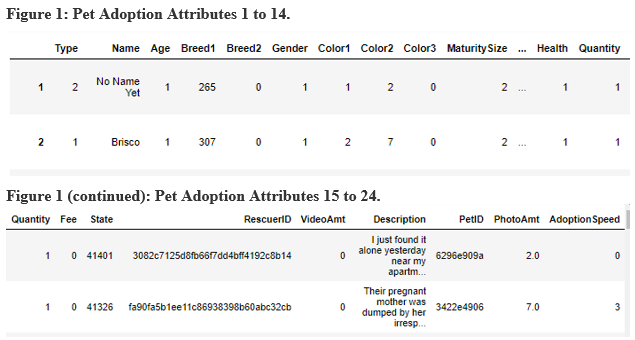
## **1.3 Dataset Summary**

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.

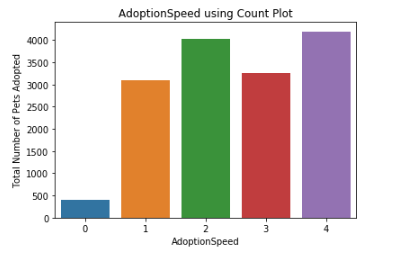
The training dataset will only be used for this project to meet the business objectives. The Pet Adoption training dataset contains a total of 24 attributes and 14,993 rows of data.The list of all 24 attributes can be seen in Figure 1. The PetAdoption dataset had three types of data classifications listed 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.

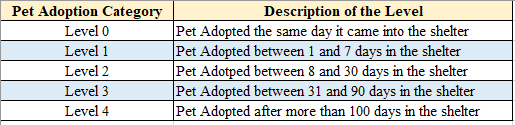
The **AdoptionSpeed** attribute is an ordinal variable that shows the length of time an animal stays in the shelter and is the target variable. The **AdoptionSpeed** variable has a total of five levels based on the length of time an animal stayed in the shelter and the description of the levels can be 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.**

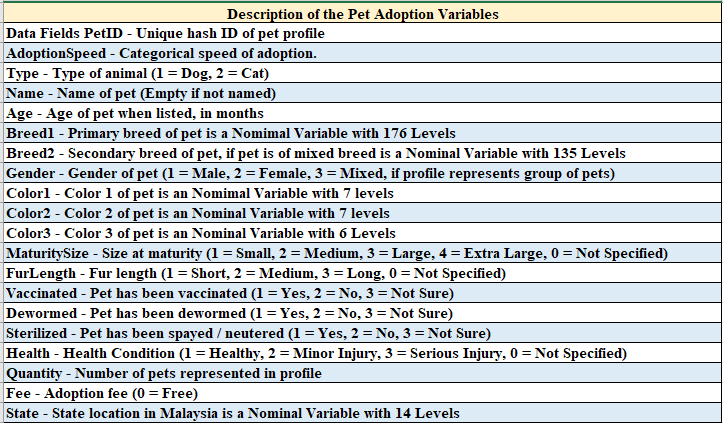


**Figure 2: Count and description of the AdoptionSpeed Variable.**

****



**Figure 3: Pet Adoption Variable Description**



# **2. Data Wrangling**

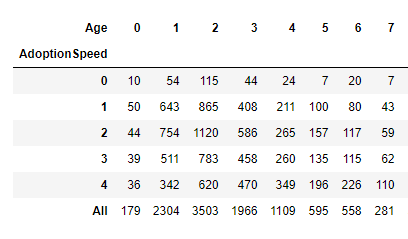
## **2.1 Missing Data Points**

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 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**



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.

**Figure 5: Boxplots for the Age and Log10 Transformed Age\_t variable**

|  |  |
| --- | --- |
| **Boxplot for Age Variable** | **Boxplot for Log Transformed Age\_t Variable** |

Breeds 266 and 307 remained as one category and then the remaining values were placed into two groups. All breeds that were less than 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.

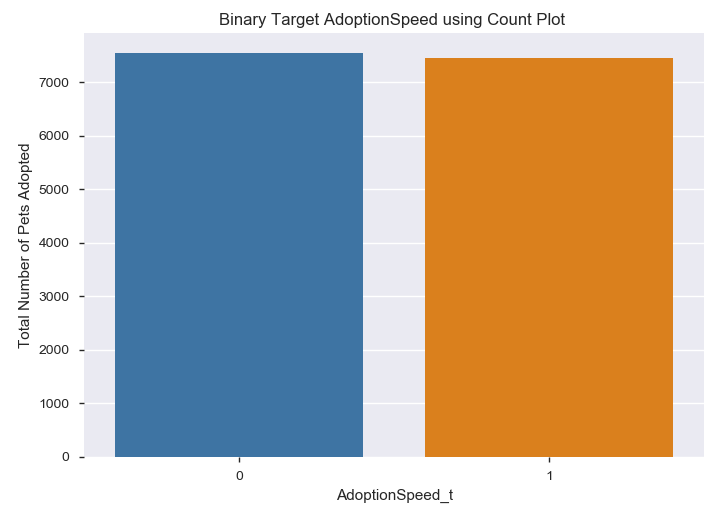
**Figure 6: Boxplots for the Breed1 and the binned Breed1\_t variable/**

|  |  |
| --- | --- |
| **Boxplot for Breed1 Variable** | **Boxplot for 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 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 at least 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 with only 1,422 for two pets listed at a time. These numbers decreased when two or more pets were included for the listing for these shelter pets. The **Quantity** variable will bin all animals that had 2 or more animals for a listing to be used for the optimized model and can be seen in Figure 7 below.

**Figure 8: Binned Results for the State\_t and Quantity\_t Variables.**

|  |  |
| --- | --- |
| Count Plot for Binned State\_t Variable | Count Plot for Binned Quantity\_t Variable |

# **3. Exploratory Data Analysis (EDA)**

## **3.1 Questions of Interest (4 Figures)**

There was a total of four variables that were tested to determine if there were no significant differences between the attributes 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 alternative hypothesis can be rejected at a 0.05 p Value and the output from the statistical results can be seen in Figures 9 to 12 below.

Figure 9

|  |  |
| --- | --- |
| Statistical Results for the Gender Variable | Statistical Results for the Type Variable |
|  |  |

Figure 10

|  |  |
| --- | --- |
|  |  |

Figure 11

|  |  |
| --- | --- |
| Statistical Output for the Sterilized Variable | Count Plot for Vaccinated Variable |
|  |  |

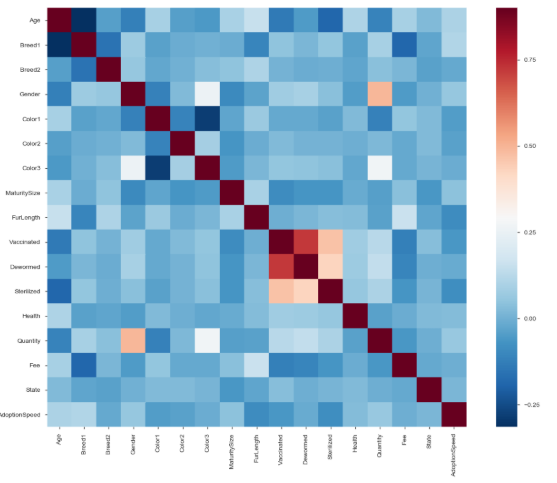
Figure 12

|  |  |
| --- | --- |
| Count plot for Sterilized Variable | Count Plot for Vaccinated Variable |
|  |  |

## **3.2 Correlation Heatmap**

The Correlation Heatmap Matrix can be seen in Figure 13 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.

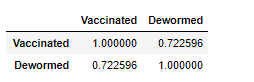
**Figure 13: Heatmap for all 24 Attributes of the Pet Adoption Data Set.**



## **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 14. 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.

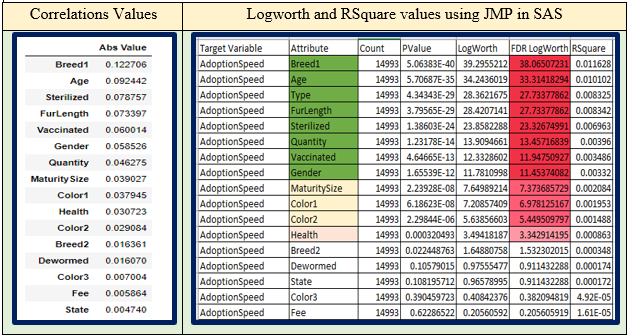
**Figure 14: Pearson Correlation Values**



## **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 13).

**Figure 15: Correlations Plot Values and Logworth Values from JMP.**



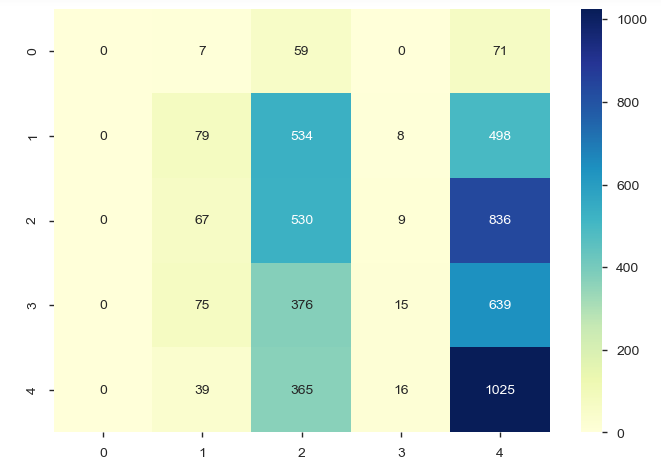
# **4. Preliminary Model Testing**

The baseline model will use all 17 original numerical input variables to get a baseline result for the five-ordinal level of the **AdoptionSpeed**. The standard Logistic Regression model will be used with C = 1 and a penalty of l1 for the model. The data was split 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 binary target variable was used for this Logistic model.

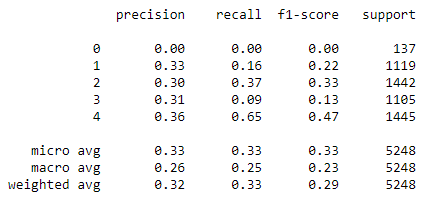
## **4.1 Baseline Logistic Regression with all Attributes**

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 16. 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 17). 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 16: Confusion Matrix for the baseline model.



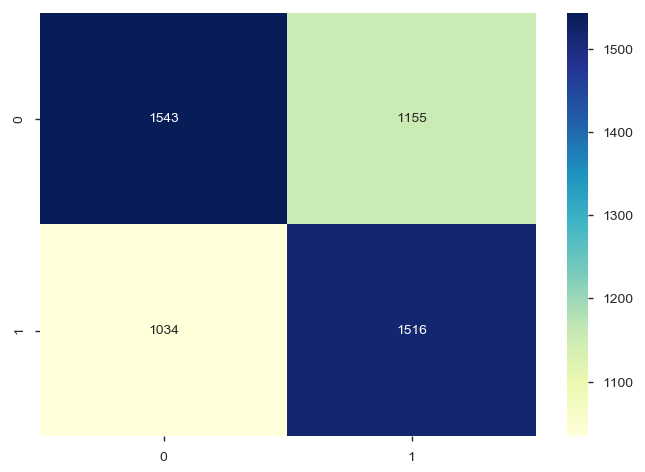
**Figure 17:** **Classification Report for the baseline model**



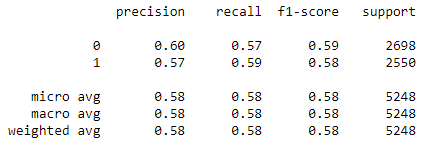
## **4.2 Logistic Regression for the optimized model**

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 18 and 19. 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 18: Classification Matrix for the Optimized Model**



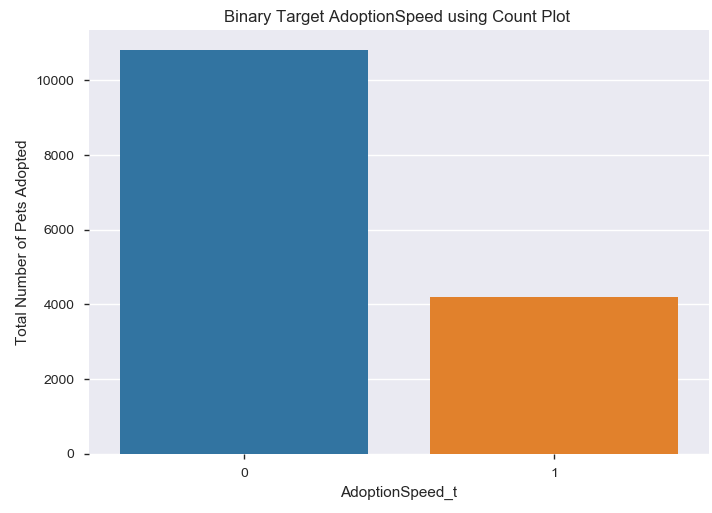
**Figure 19: Classification Report for the optimized model**



## **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 model and create graphs for the Lifespans for select attributes. The top eight attributes used for the optimized model 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 20). There was a total of 4,197 animals that remained in the shelter for more than 100 days and is population of interest in the lifeline study. The goal is to have lower longevity value for survival ship in the model and basically be adopted form the shelter. The Age attribute was selected for the duration column and the transformed AdoptionSpeed was selected for the event column.

Figure 20: Transformed AdoptionSpeed 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 and Vaccinated was the only variable that did not have a significant affect with a p value of 0.64 (Figure 21). The FurLength attribute had the only negative impact on 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: Cox PHFitter Summary Statement for select inputs.

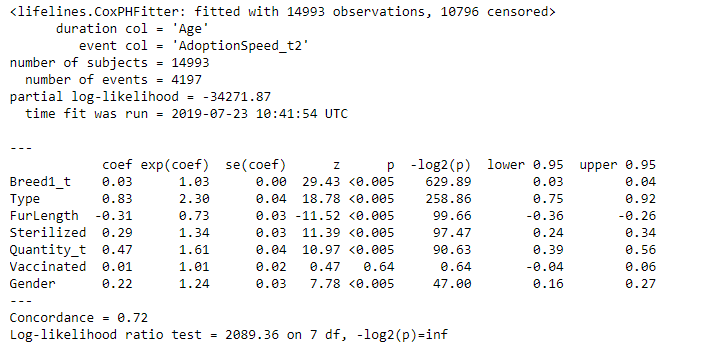
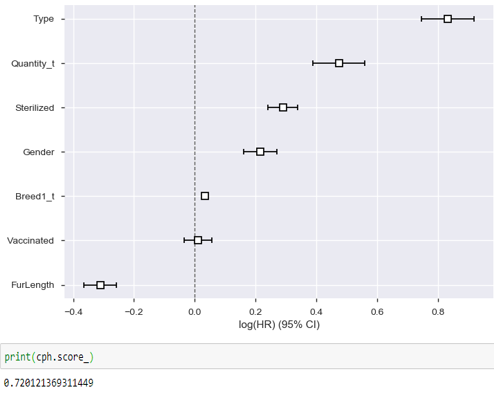
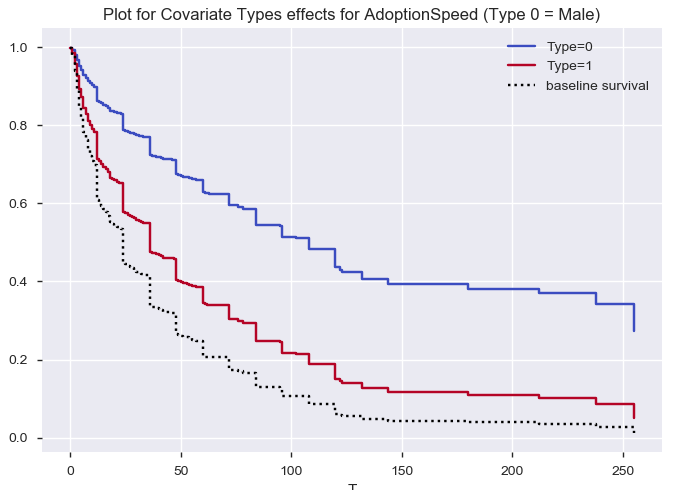


Figure 22: Plotting the Coefficients for the Cox PHFitter Model.



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 23. 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 cats for survival at the age of 125.

Figure 23: Lifespan for the Type of Pets for AdoptonSpeed

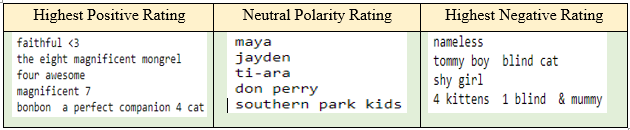


# **5.0 Text Analysis for Pet Names**

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 use 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 like adoption or new home. There are also cultural differences in the names because these Adopted pets are from Malaysia and a few words like” Mimi” for 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 24 Sentiment Polarity Rating for the top words for Positive, Neutral and Negative Values.**



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 for the 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 25 Polarity Score for Name**

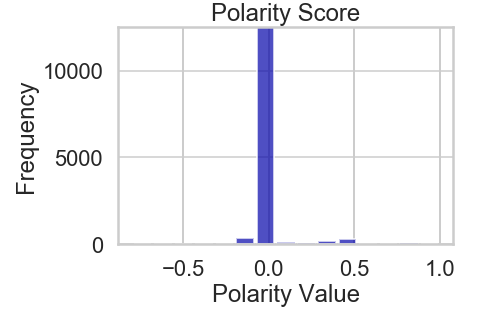
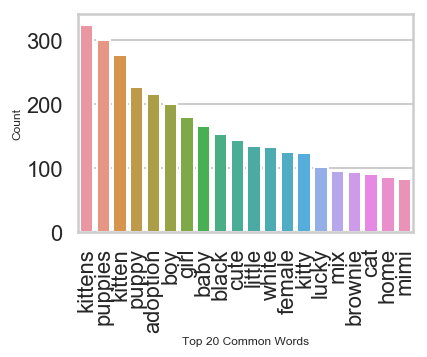


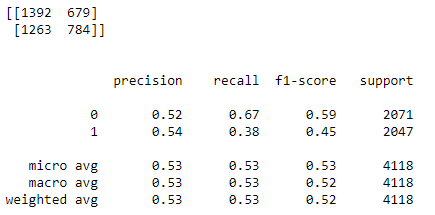
Figure 26 Top 20 Most Common Names for Pet Adoption



A preliminary test was competed for the text analysis with the **Name** and **AdoptionSpeed** using the Naïve Bayes model. The standard preprocessing steps were performed (remove punctuation, stopwords and cover to lower case) and the words were vectorized with the bow transformer approach in Python. The NB model tested the Name attribute with the five ordinal and 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 and the Logistic Regression for the Pet Adoption Model are similar in values and both are much lower and are currently weak predictors for the target variable.

Figure 27: Naive Bayes Model for testing Name Attribute with

a Binomial Target Variable Classification Report.



# **6.0 Conclusion**

# **7.0 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.