Springboard Data Analytics

January 2019 Cohort

Capstone I Project

05Feb2019

John Parsons

**Project Summary**

The objectives of this paper are to define the parameters of the first capstone project for the Data Analytics Career Track at Springboard. This paper will discuss the following items; what is the problem to be solved for the PetAdoption Kaggle Challenge, who is the client in this study and why do they care about the study, what is the type of data being used for this study, how will this study reduce adoption time for animals in shelters, minimize euthanizations and how will the results be presented to the client.

**What is the problem you want to solve?**

The **PetFinder** Data set is an existing Kaggle competition that will be used to pick the speed at which a pet is adopted based on 24 attributes of varying datatypes. The algorithms generated from the competition will be used to guide shelters and rescuers find homes for abandoned pets and minimize the number of euthanizations on pets that cannot be adopted.

The business problem is reducing the shelters cost by reducing the time pets stay in the shelters by using machine learning algorithms. The cost varies between shelters, but it is around 20 dollars per day to shelter a pet and 200 dollars for the initial veterinarian bill (deworming, vaccinations, neutering and spaying) in the United States. The Histogram generated in R for Figure 1 shows the distribution of animals over five ordinal target variables. There are only 410 animals that were adopted on the same day for target value of 0 and 3,090 animals were adopted out between 8 and 30 days (Target = 1). The biggest concern is the 4,197 animals that stayed in the shelters for more than 100 days for the target value of 4. Every day these pets stay in the shelters for more than 100 days, will cost the shelters around 83,000 per day. The goal is to significantly reduce the time these animals stay in the shelters and adopt them out before the 100-day period has lapsed for each animal.

The second business objective is to have more positive consumer ratings for the Nestle Company which has purchased the PetFinder App. Nestles model is “pets and people are better together” and by finding those champion models that will successfully adopt out pets and minimize the number of euthanizations for unadoptable pets, is a win-win solution for the pets, shelters and Nestle Company.

**Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?**

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. The champion model that is generated to decrease shelter time for pets 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. The last major benefit is finding homes for over 250,000 pets that have yet to be adopted and find ways to lower the number of adoptions that occur worldwide and make this be a successful campaign for all parties.

**What data are you using? How will you acquire the data?**

The PetFinder.my Adoption Dataset is coming from the Kaggle competition. This data set contains a total of three files to be used for the **AdoptionSpeed** at which an animal is adopted from the shelter (Figure 2). The dataset contains three files for determining the rate at which an animal is adopted from the shelter. The first file is a metadata set that contains a total of 24 attributes and 14,993 rows of data. There are no missing variables in the meta dataset as shown in Figure 3 from the **AmeliaView** in R but the Python using the ebola statement shows missing values for the Name and Description Attributes. The summary, description and missing data results from R can be seen in the appendix (Figures 4, 5 and 6).

The **PetAdoption** Speed attribute will be the Target variable and has a total of five levels or categories. The remaining 23 attributes are the Input variables to be used to develop a model to predict adoption speed of a pet. Figure 7 shows the output in Python of the datatypes for all 24 attributes in the dataset. Almost all attributes are categorized as an Int64 except for the **Name, RescuerID, Description** and **PetID** which are an Object and **PhotoAmt** which is a Float64. The **PetId** and **RescuerID** are unique identifiers and will not be used in the analysis and the name and description will be used in the Text Analysis section of the study. There are a few additional Excel CSV Files used to describe the types of colors and breeds of these animals found in the training data set.

The second two files are unsupervised files and contains a set of pictures (jpg files) and text files (Jason files). These two files are an addition to the original training file and can be used to improve the model’s performance. The goal of this study is to use the training dataset to develop the best champion model to reduce shelter time for the pets based on the most significant attributes. The text files and photos will be analyzed separately to do sentiment analysis on the most favorable words used in this campaign and select the photos that were successful in adopting out these pets.

**Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

Python Notebook and SQL will be used for data preparation, cleaning and exploratory data analysis of the metadata set. Data wrangling includes smoothing outliers, removing redundant and irrelevant input variables, minimizing the skewness and standard deviation from the attributes. This data set only contains missing values in the name and description of the data set and the other 22 attributes have no missing values.

Data analysts need to minimize the number of attributes used for the models to reduce dimension space and help model performance. The p-values, logarithmic worth or other factors can be used to determine the best attributes to be used in machine learning models. These values were initially determined in JMP SAS and the results can be seen in Figure 8. A total of eight attributes have been selected to be used as the input variables for the machine learning algorithms to predict adoption speed. The other attributes may be added to increase the performance of the model after the champion model has been selected.

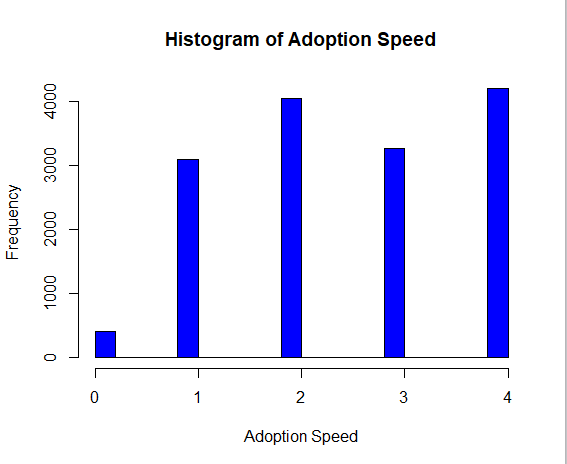
The first eight input variables will be used to develop several models (Decision Trees, Logistic Regression, Neural Nets, Random Forests and Ensemble Models) to determine the best approach for reducing Shelter time for these pets. The data will be divided into a training and validation data set (65:35 split). The validation data set will be used to determine the model’s performance and the models that has the lowest misclassification rates will be selected as the champion model.

**What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

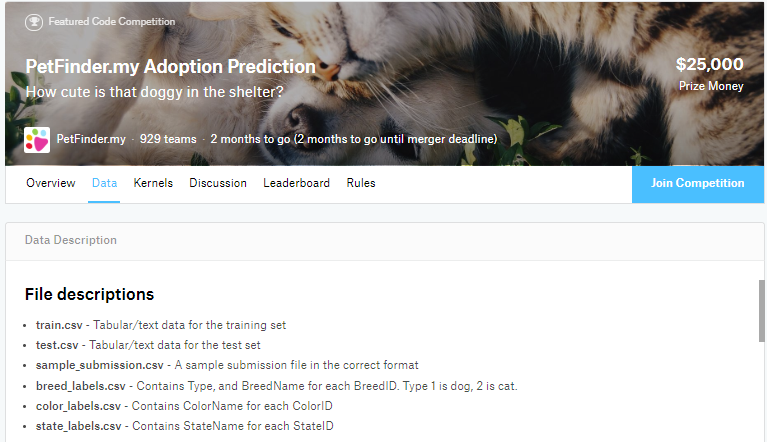
The deliverables from this project will contain the PetAdoption Final Report that will contain the following items; Executive Summary, Introduction, Data Set Description, Project Objectives, Data Preparation, Data Analysis Methods, Results and Future Recommendations based on this study. The Python code and SQL will be submitted in a text and or HTML File for this project. All files will be uploaded to GitHub to present the findings from this study.

**Appendix**

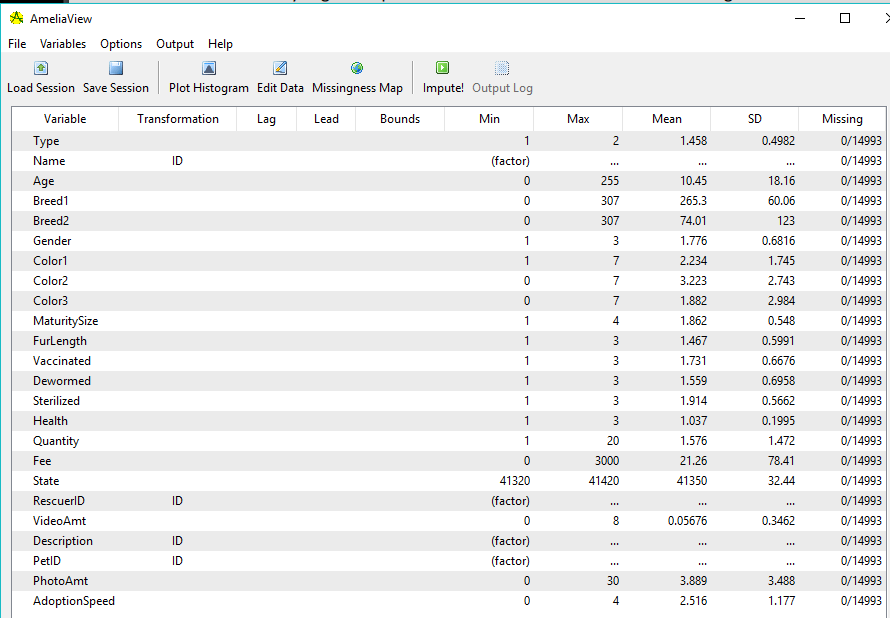
**Figure 1: Histogram of Pet Adoption Speed.**



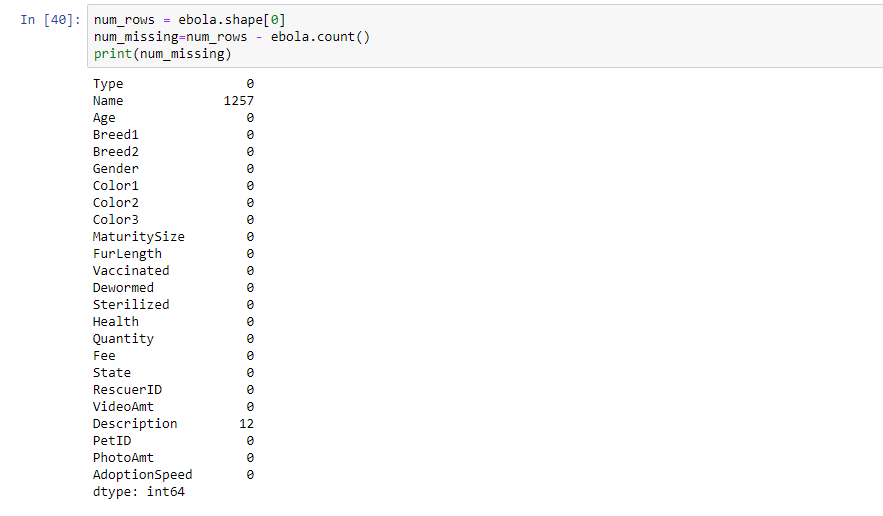
**Figure 2: Snapshot of the datafiles for the PetFinder.my Adoption Prediction Challenge.**



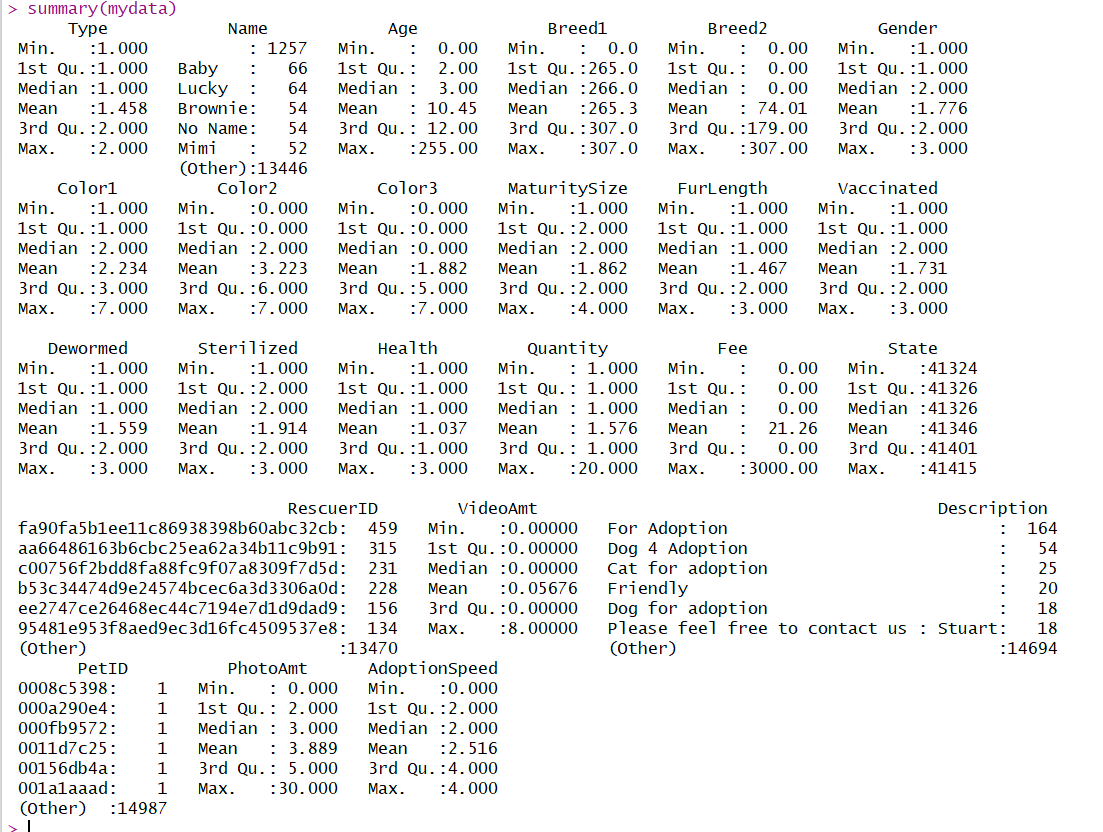
**Figure 3: Amelia II View statement in R showing there are no missing values.**



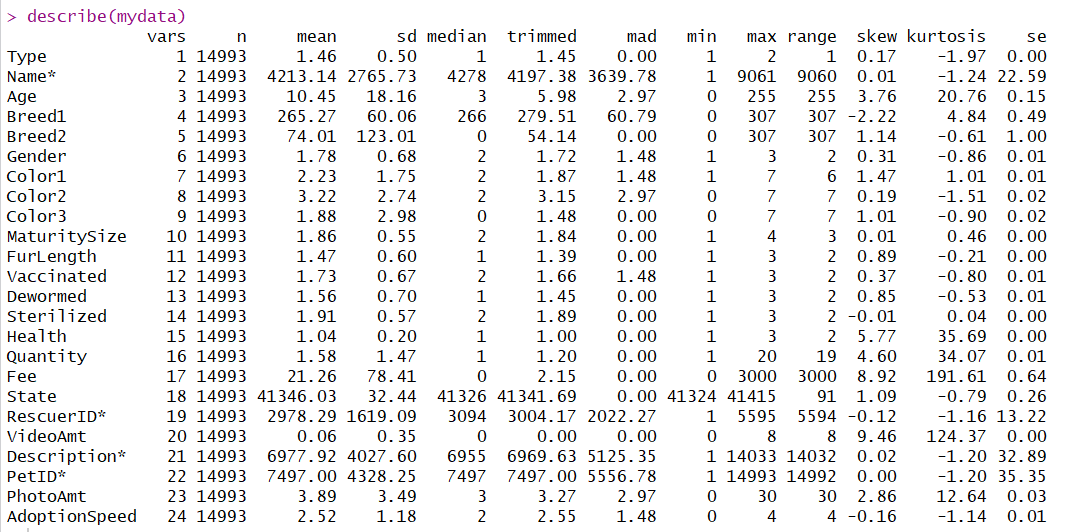
**Figure 4: Python Output using ebola to determine the number of missing attributes.**



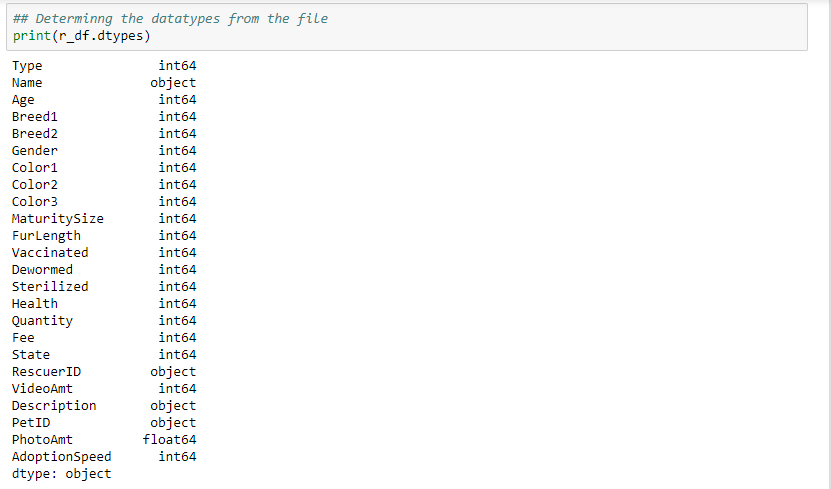
**Figure 5: Summary statement in R Studio from the metadata set.**



**Figure 6: Describe statement in R for the variable statistics.**



**Figure 7: Datatypes in Python using dtypes command.**



**Figure 8: P-Value, Logworth and RSquare values of the Input variables in JMP**

