Springboard Data Analytics Course

Model Development for Capstone I

13 June 2019

John L. Parsons

**Objectives**

This progress report will give a brief overview of the Pet Adoption data set, the data wrangling steps used for these variables, the attributes that will be used for the baseline model and the model optimization steps. There are three phases for the model development section and the first one will use the Decision Tree and Logistic Regression to find the baseline model with the best accuracy score. The second phase will contain a total of eight and ten models that will use eight unoptimized models and ten models with ten Kfold validations to optimize the model’s performance. The best models will be used for an Ensemble Model for this section. The last phase for this study is to conduct a text analysis for the **Description** attribute to determine the polarity score and standard word count information to better understand how each animal was characterized in this study. The model results can be found in the Appendix of this paper and the description of the models will be discussed for each section.

**Model Development**.

The pet adoption file contains a total of 24 attributes that were included in the dataset. All non-numeric attributes were removed from the **AdoptionSpeed** to create the machine learning models. The variables that were removed are; **Name**, **Description**, **PetID**, **PhotoAmt**, **VideoAmt** and **RescuerID**. The remaining 18 attributes can be seen in Figure 1 and this includes the number of missing values, categorical levels and name of the attribute. The **AdoptionSpeed** is the target variable and is an ordinal variable with a total of five levels. Level 0 represents animals being adopted the same day and ends with level 4 which is for all animals that has been in the shelter for more than 100 days. The 17 input variables ranged from a binary variable such as **Type** to a nominal variable with 117 unique values as shown in **Breed1**.

The data cleaning steps for these 18 attributes were minimal because all variables were numeric, no characters and had no missing values in the data set. A total of 12 of the 18 variables had two to seven levels per category recorded for this dataset. The **State**, **Quantity**, **Fee**, **Age**, **Breed1** and **Breed2** had more than seven unique levels ranging from 14 to 117.

The dataset contains a total of 14,993 rows or samples and will be divided into a training and validation (test) data set. The training data set will contain 65% of the total training data and the remaining 35% will be used for the test set. A small subsample containing 10% of the original training sample will be used to test the model optimization steps before the entire set was tested for the final results. All numeric attributes for the **PetAdoption** file will be used to find a baseline model and then the best attributes will be selected to improve the baseline models performance.

The machine learning model development stage can be divided into three stages. The first stage contains the standard **Decision Tree Classifier**, **Logistic Models** and parameter optimization methods for these models (Gini, entropy, best C and best penalty). The goal is to get the baseline value to determine how well the model fits the data. The confusion matrix, precision, recall and F1-score was generated from these models to determine the best baseline values for the model.

The second model development phase will test two groups of models. The first group contains eight unoptimized models (eight model method) to determine the accuracy test results for the data set. These models are **Random Forest, Naïve Bayes, Decision Trees, Logistic Regression, KNN, Support Vector Machines, Perceptron and Stochastic Gradient Descent**. The training and test accuracy scores will be calculated to see if this improves the accuracy score from the baseline model. Then a group of 10 models (ten model method) will be used for Kfold Stratified Cross Validation with a Kfold of 10 for model improvement. These models are **Support Vector Machine (SVC), Decision Tree, AdaBoost, Random Forest, Extra-Trees Classifier, Gradient Boosting, Multiple Layer Perceptron, KNN Nearest Neighbors, Logistic Regression** and **Liner Discriminant Analysis**. The accuracy score, learning curves and Relative Variable Importance Features will be calculated from the top models based on accuracy score and used for further model development. Finally, an Ensemble Model will be created using the top models to improve the model’s performance.

The third phase is the text analysis of the **Description** attribute for the pets in the shelter. These text files will be cleaned, stop words removed, whitespace will be removed and capital letters will be transformed to lower case letters. The document will then be processed to determine the sentiment analysis for the pet description. The text analysis will also count the top common words with CountVenctor, top Bigram and Trigram words found in the document and Term Frequency and Inverse Document Frequency for the Pet Adoption file. The polarity scores, word length and frequency will be added to the eight-variable model to improve the model’s accuracy score.

The attributes that will be used for these models will first contain all 17 numeric variables for the input values and the five-level ordinal variable **AdoptionSpeed** as the target variable. The 17-attribute data set will be tested in both model phases of this study to determine the best test accuracy score that will become the baseline estimate for this dataset. Then a total of the eight top attributes based on variable importance, Pearson Correlation and Logworth Values have been selected to improve the model performance. The eight input variables are; **Age, Breed1, Furlength, Quantity, Vaccinated, Gender, MaturitySize and Sterilized**. The target variable will be merged from the five-level ordinal variable to a binary variable. The target variable will have levels 0, 1 and 2 binned into a 0-level and any animal that has been in a kennel for less than 30 days will be in this category. 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 in the shelter.

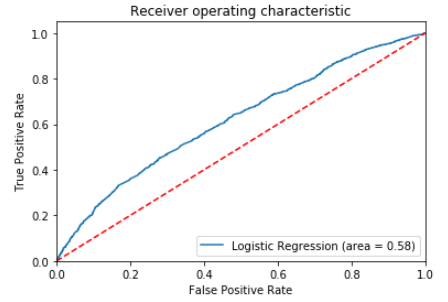
**Model Results**

The Logistic and Decision Tree Classifier models both had a Test Accuracy Score of 33% for the 17-attribute data set with no transformations being made to these variables. The results of the optimized Logistic Regression Model can be found in Figure 2 and the Confusion Matrix shows the model had a very hard time classifying the **AdoptionSpeed** Categories 0 and 3 for this data set and Category 4 had the best results in terms of the precision, recall and F1 values.

The results from the 17-attribute data set for the eight-model method can be seen in Figure 3 and the top two results were from the Random Forest and Naïve Bayes. The test accuracy results ranged from 38.5% for the Random Forest and 33.6% for the Naïve Bayes. These results were slightly better than the previous two models with an accuracy score of 33%. The Ten model Cross Validation for in Figure 4 shows Gradient Boosting and then the Random Forest with the best mean accuracy score with values between 35 and 40%. The ten-model Cross-Fold optimization test performance results increased performance slightly to 40% for Random Forest Classifier, Logistic Regression and KNN models. The Random Forest Classifier had the best accuracy score at 40.2%. The top four models were used in the Ensemble Model and these are Logistic Regression, Random Forest, KNN and Gradient Boosting Classifier to improve the accuracy score. The accuracy score was only 37.8% and did not do as well as the Random Forest Model. The results from the ten model optimization steps can be seen in Figures 5 to 7.

The eight-attribute data set (this contains eight input variables and a binary target variable) was tested with the two model development phases to see if this can be a better predictor of **AdoptionSpeed**. The model accuracy score was 61% for the Decision Tree and 58% for the Logistic Model and the results can be seen in Figure 8 and 9. The Precision, Accuracy and F1 scores were near the midpoint of 50% for all three categories of the Decision Tree and between 57 and 59% for the Logistic Model. This is an improvement compared to the 17-attribute dataset for the baseline model, but it is still a weak performing model as shown in the ROC Curve in Figure 1A below.

**Figure 1A: ROC Curve for the Logistic Regression Model.**



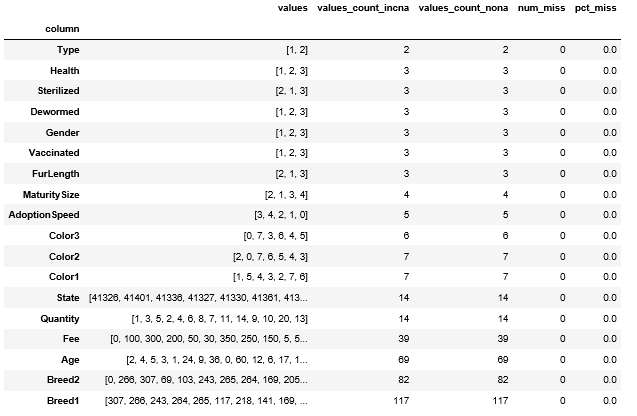
The eight-variable dataset was then tested with the 10 models and the results can be seen in Figures 10 to 12. The preliminary results had three models with accuracy scores greater than 60% and these are the SVC, Gradient Boosting and Multi-Layer Perceptron models and the other seven were slightly lower than the 60%. The best results using the KFolds was from the Gradient Boosting Model and this had an accuracy score of 64.2% and is a slight improvement over the previous models. The top four models were used in the Ensemble Model and the test accuracy score dropped to a value of 61.9% and was not as efficient as the Gradient Boosting Model. The variable importance can be seen in Figure 12 and the **Age**, **Breed1** and **Vaccination** were the top three important attributes and the remaining four variables were much lower and consistent in their score as shown in the graphs.

**Text Analysis**

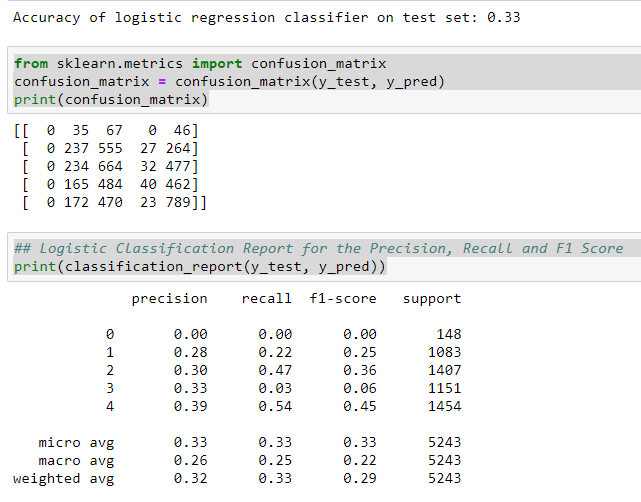
The Text Analysis was calculated on the **Description** variable from the Pet Adoption file. The polarity scores, most common words, Bigrams, Trigrams and Term Frequency and Inverse Term Frequency per document or individual animal results can be seen in Figures 13 to 16 in the Appendix. The polarity score had a median value around 0.25 and was slightly more positive with some negative tone to the description. One problem with the description category is many words are not from the English Language and this may have affected his rating. The top 20 Bigram graph shows common two-word phrases that contain ‘**home’, ‘loving’, ‘months old’, new home and ‘adopt**’. The polarity score, Length of Review and Word Count was added to the eight-variable model to see if the model’s accuracy score could be improved and this actually had no effect on the model’s performance as seen in Figure 17. The test accuracy score is still 58% for the optimized Logistic Model and has similar values for the Precision, Recall and F1 scores (values ranging from 57 to 59) and this did not improve the model’s performance.

**APPENDIX**

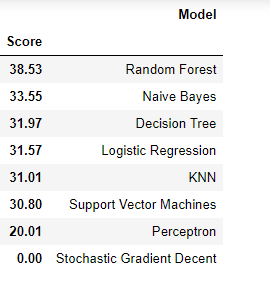
**Figure 1: Attribute levels, missing values and unique quantity count for the dataset.**



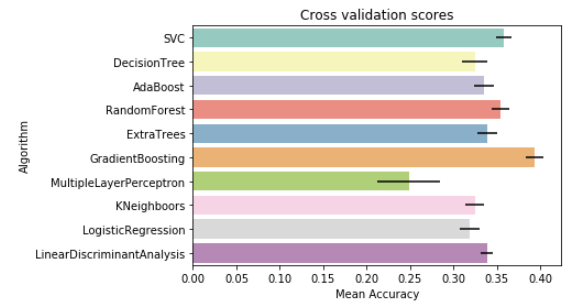
**Figure 2 Model Results for Logistic Regressing for all 17 input attributes.**



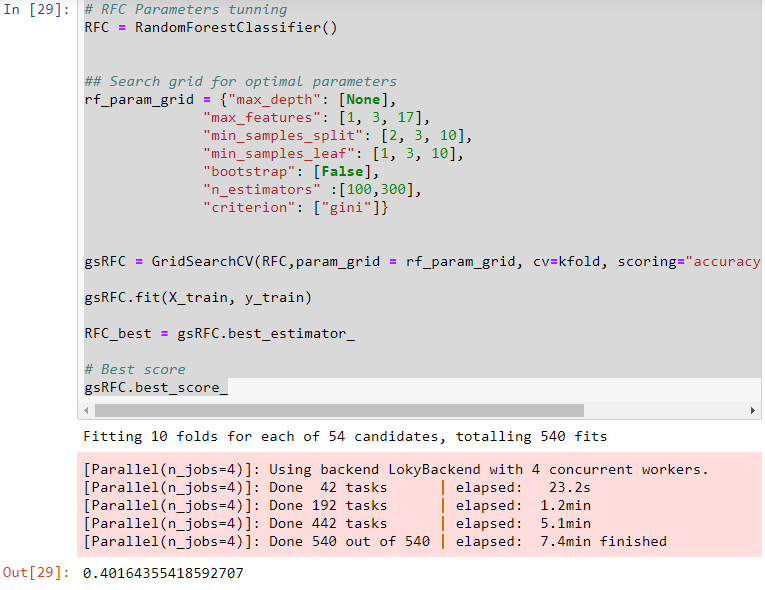
**Figure 3: Eight Model Test Accuracy Results.**



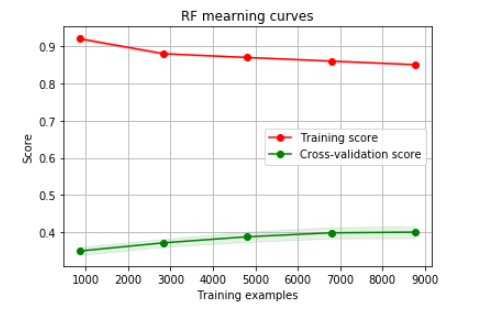
**Figure 4: Ten Model Cross Validation Test Accuracy Score Results.**



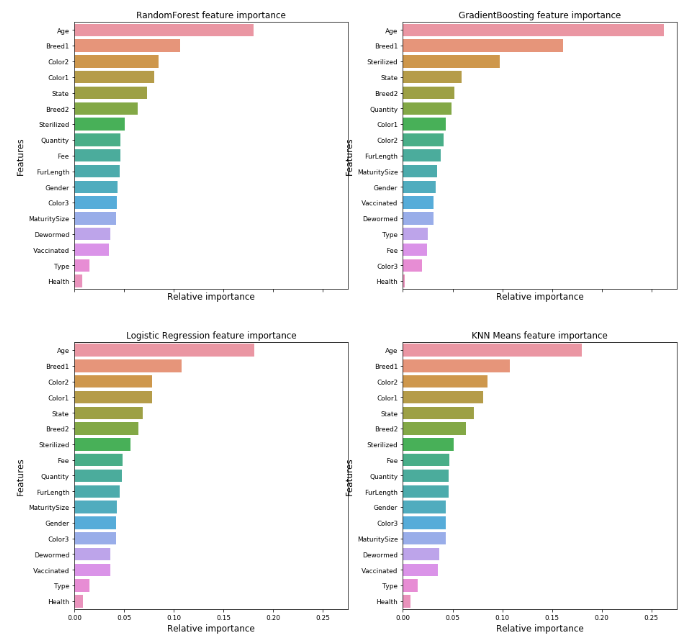
**Figure 5: Random Forest Classifier Results.**



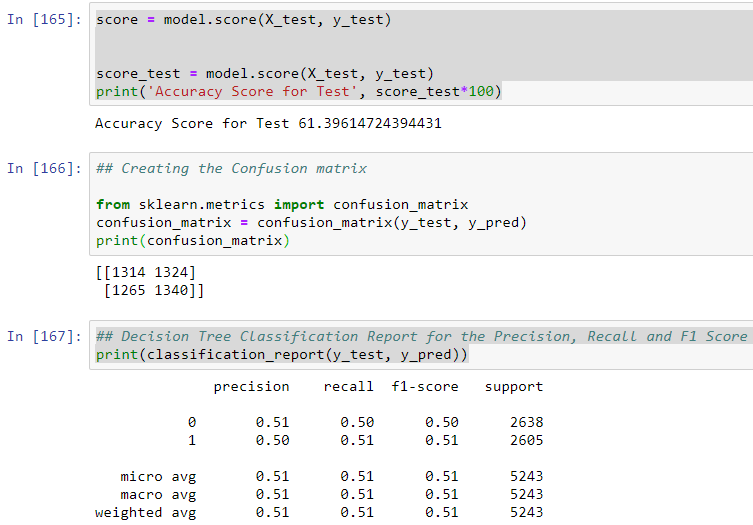
**Figure 6: Random Forest Classifier Learning Curve Results.**



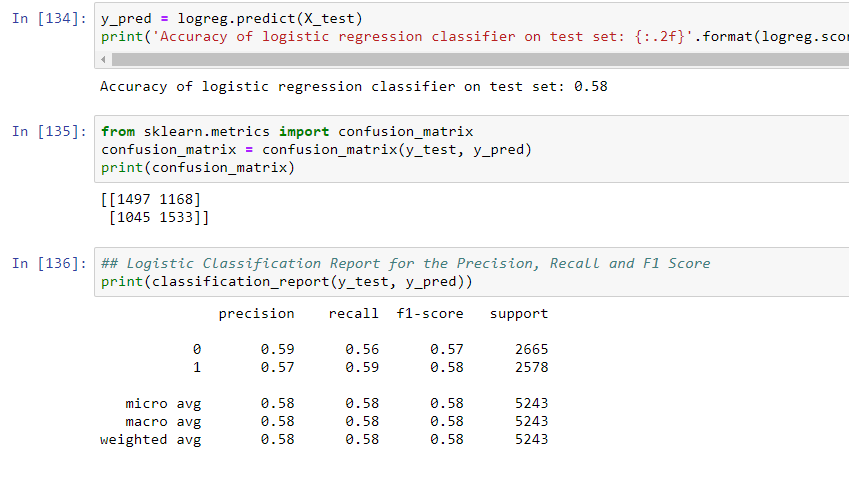
**Figure 7: Feature Importance Values for the top four models.**



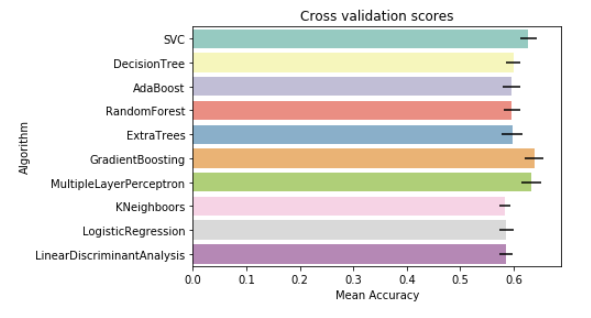
**Figure 8: Decision Tree with the eight-variable model.**



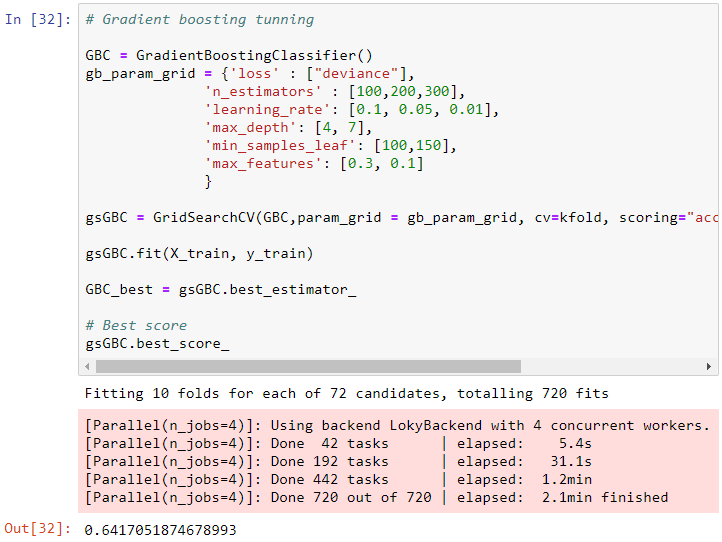
**Figure 9: Logistic Regression with the eight variable model results.**



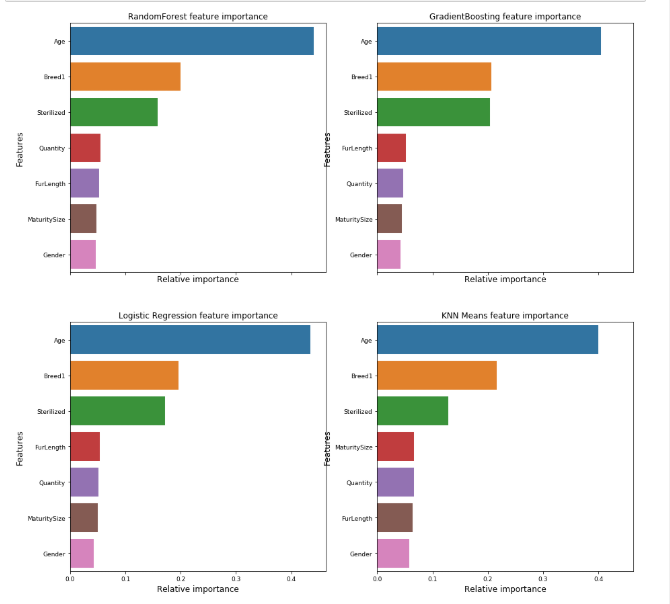
**Figure 10: 10 Model Mean Accuracy Scores for the eight-variable model.**



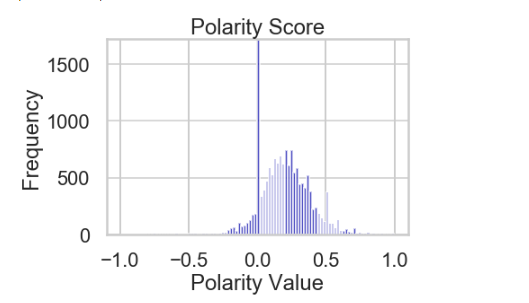
**Figure 11: Model Optimization for the Gradient Boosting Model.**

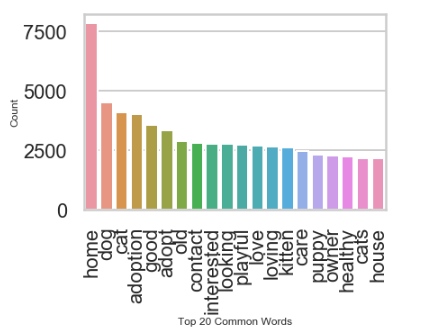


**Figure 12: Variable Importance for the Top Four Models.**

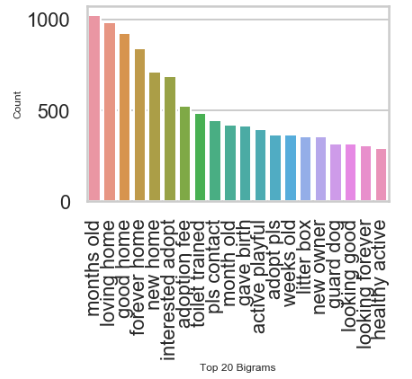


**Figure 13: Polarity Score for the Description Attribute.**

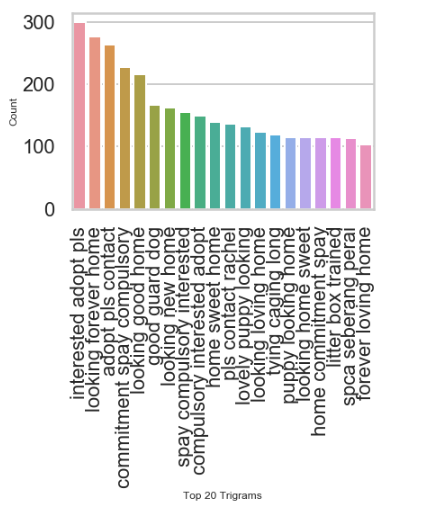


**Figure 14: Top 20 Common Words for Description Attribute**.

**Figure 15: Top 20 Bigrams for the Description Attribute.**



**Figure 16: Top 20 Bigrams for the Description Attribute.**



**Figure 17: Logistic Model for the Eight Attribute Model with Polarity, Word Length and Word Count.**

