Final Project: Animal Rescue Returns Classification

**Naïve Project Team**  | Rachel Kelley (gtID: 903749193) and Justin Schulberg (gtID: 903737135)

Breakdown of Roles

For the most part, the tasks in the project were equally-shared between the two team members, Rachel and Justin. In particular:

|  |  |
| --- | --- |
| Task | Assignee |
| Requirements Gathering | Rachel (Support) & Justin (Support) |
| Data Preparation | Justin (Lead) |
| Data Cleaning | Rachel (Lead) & Justin (Support) |
| Classification Models + Analysis | Rachel (Support) & Justin (Support) |

For a more thorough breakdown of the tasks covered by each team member, feel free to access the [*Issues* page of the GitHub project](https://github.com/jschulberg/Dog-Returns/issues), where various tasks are assigned to the team member who completed them.

Problem Statement

For this final project, the Naïve Project Team will be analyzing a dataset from [a local Animal Rescue](https://www.luckydoganimalrescue.org/about), a nonprofit dedicated to rescuing homeless, neglected, and abandoned animals from euthanasia in kill shelters and getting them adopted into their forever homes. The nonprofit educates the community and all pet parents on responsible pet parenting, including the importance of spay/neuter, obedience training, and good nutrition.

Successful adoptions of cats and dogs depend on a variety of factors, both on the parts of the pet and of the family. The nonprofit helps bridge the gap, building strong relationships between the two. The nonprofit rescues hundreds of animals every year, provides them with loving temporary care, and finds them well-matched, carefully screened forever homes.

One of the struggles that the nonprofit runs into is pets being returned after adoption. Even though they adopt out ~2000 dogs per year, about 10% get returned for a variety of factors. The nonprofit assiduously tracks information on all their adoptions and returns. **In this project,** **the** **Naïve Project Team will use various analytical methods learned in Computational Data Analysis to help the nonprofit predict whether or not a dog that they adopt out will be returned**.

Data Source

The project will be overseen by the *Program Manager for Volunteers and Data Integrity* at the nonprofit. She will provide the data needed for this analysis to the Naïve Project Team. Data will be provided across various spreadsheets, detailing the adoptions and returns per year (i.e. Adoptions 2020, Returns 2018, etc.) from the nonprofit, both of which contain records over the past decade, and can be easily linked by a unique ID field for each dog represented in each dataset. The team has worked to programmatically concatenate the various adoption and returns spreadsheets from various years, which unfortunately are not always in the same format and sometimes have different column headers, into two main data sources:

* **Dog List** | A list of every dog that has been adopted out by the nonprofit over the past 10 years. This list includes a variety of features describing each dog, where each row corresponds to a dog being adopted out, and each column represents a different attribute related to that dog. This dataset is composed of 10 spreadsheets, one for every year over the past decade. Each spreadsheet has records of ~2000 adoptions per year. The attributes in this dataset are as follows:

|  |  |
| --- | --- |
| **Field** | **Description** |
| **Dog Name** | Name of the adopted dog |
| **ID** | Unique ID corresponding to that dog |
| **Link** | Link to dog’s profile on the nonprofit site |
| **Foster/Boarding** | Type of adoption for the dog (short vs. long-term) |
| **Sex** | Gender of dog |
| **Age** | Estimated age (sometimes a range if actual age not known) of dog at time of adoption |
| **Weight** | Estimated weight of dog at time of adoption |
| **Breed Mixes** | Type of dog |
| **Color** | Color of dog |
| **Behavioral Notes** | Free text field describing the behavior of the dog. There’s some consistency in entries depending on the individual entering this field, along with some key words to describe the dogs behavior around others |
| **Dogs in Home** | Number of other dogs in adoptee’s house, if any at all |
| **Cats in Home** | Number of cats in adoptee’s house, if any at all |
| **Kids** | Number of kids in adoptee’s house, if any at all |
| **BS/W** | Indicator for whether or not the dog is part pitbull or other ‘bully’ breed |
| **Medical Notes** | Free text field describing any health conditions of the dog, if any exist at all |
| **Transport Date** | Date of adoption |

* **Returns List** | A list of every dog that’s been returned *after being adopted out* by the nonprofit. This list also includes a variety of features describing each dog, where each row corresponds to a dog being adopted out, and each column represents a different attribute related to that dog. This dataset is composed of 10 spreadsheets, one for every year over the past decade. Each spreadsheet has records of ~200 returns per year.

|  |  |
| --- | --- |
| **Field** | **Description** |
| **Dog Name** | Name of the adopted dog |
| **ID** | Unique ID corresponding to that dog |
| **Dog Info** | Free text field with general information on the dog |
| **Reason for Return** | Free text field describing the reason the dog was returned by the adopters. There’s some consistency in this field, but it generally depends on the person entering the information into the spreadsheet |
| **Behavior with Dogs** | Free text field describing the dog’s general behavior around other dogs |
| **Behavior with Kids** | Free text field describing the dog’s general behavior around kids |
| **Behavior with Cats** | Free text field describing the dog’s general behavior around cats |
| **Energy Level** | General energy of the dog |
| **Socialization/Daycare** | Categorical field for whether the dog was in a daytime program for dogs |
| **Vetting** | Categorical field for whether the dog was being taken to the vet |
| **Date of Adoption** | Date dog was adopted |
| **Previous Return?** | Boolean for whether the dog has been returned in the past |
| **Previous Return Info** | Free text field describing why the dog had been returned in the past |
| **Date of Return** | Date dog was returned |
| **Type** | Boolean for dog vs. puppy |

Data Preparation/Cleaning

To begin, we processed, prepared, and cleaned the data for analysis. Currently, all data is kept by the nonprofit in different spreadsheets by year (i.e. adoptions in 2021 are in the ‘Dog List 2021’; so we programmatically combined all of our datasets. On top of that, the Dog and Returns Lists are kept as two separate data sources, but share a linking variable ‘ID’ for dogs that were returned. We used a Left Join to find dogs that were returned; for any dogs which did not have a match in the join, we flagged those as ‘Not Returned’.

There were also some issues with the features provided in the data that needed to be addressed. Some of the work done to properly clean the different features is delineated below:

### [COLOR](https://github.com/jschulberg/Dog-Returns/issues/15)

We created a new column called *COLOR\_FIXED*, which is the cleaned up version of the COLOR column. After cleaning up the COLOR column with *COLOR\_FIXED*, we also made the following new features for predictive purposes:

* **multi\_color** | Variable indicating whether or not a dog has multiple colors denoted in it. This is calculated by seeing if the number of commas (',') in the COLOR\_FIXED column is greater than 1 or whether the color is denoted as 'tri-color'.
* **num\_colors** | Continuous variable counting up the number of commas (',') in the COLOR\_FIXED column.
  + *Note: This feature is highly correlated to the feature* ***multi\_color*** *since any dog which has num\_colors = 1 🡺 multi\_color = 0; conversely, any dog which has num\_colors > 1 🡺 multi\_color = 1. Though we believe that* ***num\_colors*** *will be a better predictor than* ***multi\_color*** *because it contains more information, we will test each of these features in our models* ***separately*** *to see which one ends up being more predictive.*
* **contains\_black** | Variable indicating whether 'black' appears in the COLOR\_FIXED column. This feature, in particular, is being incorporated to test the [hypothesis that black dogs are adopted less/returned more](https://en.wikipedia.org/wiki/Black_dog_syndrome) frequently than other-colored dogs.
* **contains\_white** | Variable indicating whether 'white' appears in the COLOR\_FIXED column.
* **contains\_yellow** | Variable indicating whether 'tan/yellow/golden' appear in the COLOR\_FIXED column.

### [GENDER](https://github.com/jschulberg/Dog-Returns/issues/11)

From the *GENDER* column, due to issues with the manual entry of values, we had to standardize all values as either ‘Female’ or ‘Male’, and then one-hot encoded this; 1 for ‘Male’, 0 for ‘Female.

### [BREED MIXES](https://github.com/jschulberg/Dog-Returns/issues/13)

We cleaned the *BREED MIXES* column, which denotes a dog's breed. Because there are so many combinations of different breeds for dogs, we opted to create indicator features for the most popular groupings of dogs (Lab/Retriever, Shepherd, or Other).

### [MIX](https://github.com/jschulberg/Dog-Returns/issues/14)

The column, *MIX*, denotes whether a dog is a mixed breed. We cleaned up some of the inconsistencies in the denotations of multiple breeds (i.e. the use of ‘&’, ‘and’, ‘/’, ‘with’, etc.). Also, we leveraged the *SECONDARY BREED* column to fill gaps for nulls. Lastly, we looked at the *BREED MIXES* column for a slash ‘/’ or and ‘&’ to determine if a dog is multiple (mixed) breeds. Lastly, we one-hot encoded this feature; 1 for ‘Mix’, 0 otherwise.

### [WEIGHT](https://github.com/jschulberg/Dog-Returns/issues/12)

The column, *WEIGHT*, denotes the weight, in pounds of a dog. Although this column was more consistent than others, some dogs have their weight as just a number (i.e. 60) while others have their weight along with the unit of measurement (i.e. ‘60lbs’ or ‘60 pounds’).

Analysis, Evaluation, and Final Results

After preparing the data, we will analyze the data using the various classification models that we have learned so far in class including, but not limited to, Naïve Bayes, K Nearest Neighbors, SVM, Logistic Regression, and Neural Networks. We will split the data into training, test, and validation sets, as we are using multiple models. By testing each of these methods, we hope to identify which one is most successful in classifying the adopted dogs as returned vs. not-returned.

To evaluate our models, we will look at their classification/misclassification rates using a confusion matrix, as well as their precision, recall, and F1 scores. These metrics are essential in judging the outcome of a classification model, which is why we will use it to evaluate ours. We will compare the results of our test and validation sets in order to determine which model is the most accurate. We’ll also take note of the performance efficiency of each model.

If our models are accurate, they can be integrated into the nonprofit’s systems to classify new dogs that come into the nonprofit. If a dog is classified into the return group, the nonprofit can ensure that the dog is well matched with its new adopter and reach out to the adopter to provide additional support to prevent the dog from being returned. We plan to construct this project so that the nonprofit’s data team can continue to use it, helping to place more rescued dogs in loving homes and keep them there.