Machine Learning Approach to Predictive Outcome for Sheltered Animals

Jasper Sandhu

SID: 862188200

University of California Riverside

UCR MSOL Data Science

12/04/2021

1. Abstract

Everyday there are millions of stray animals out suffering in the streets or are euthanized in animal shelters all over the world. If we can help improve any aspects of the animal shelter processes and make it easier to find homes for pets who are in need of a new family, we could reduce the number of animals that languish in shelters or euthanized due to overcrowding. With the help of the largest no kill animal shelter based out in Austin Texas and the latest machine learning algorithms we have a chance to improve our shelter processes thus increasing the survivability of stray animals and creating happiness for the animal and new found family. In this study we'll be using two supervised learning classification models with various data cleaning and feature engineering steps. The machine learning models that will be used are logistic regression and random forest classifier. The two models will be evaluated with the use of multiple evaluation metrics such as confusion matrices, and various scoring tools that includes F1-Score. Comparing the metrics for both models, we found the random forest classifier excelled with an accuracy of ~90% over logistic regressions ~68%. Ultimately the models will be able to predict the likelihood of an adoption based on the features entered by the user.

2.1 Project Overview

In the United States, over 6.5 million dogs and cats are placed in animal shelters each year. While these shelters are temporary homes for these furry critters, there's an offhand chance they could be euthanized, healthy or not if a permanent home isn't found in time. Typically about 50% of cats and dogs that are able to make it to a shelter are put to death yearly (Bradley & Rajendran, 2021). In the US there's two types of shelters: traditional and no-kill. Traditional shelters will euthanize animals depending on various circumstances while no-kill shelters will do their best to keep animals alive regardless of a overcrowding status of the facility (Sencer, 2017).

Having experience adopting two cats at separate instances from different animal shelters, with confidence I can say they've improved my health and alleviated any stressful moments while pursuing my adventure to be a data scientist. Unfortunately, not all pets are as fortunate finding a permanent home as each year approximately 1.5 million shelter animals are euthanized in the United States (670,000 dogs, and 860,000 cats) (ASPCA, 2017). While euthanasia is not the main topic for this capstone, it should be made aware of the cascading effects it has on society. Physicians who are requested to perform euthanasia may experience a conflict with their profession: the duty to preserve life on one side and the duty to relieve suffering on the other (Evenblij et al., 2019). With the help of a published data set maintained by Austin Animal Center, the largest no-kill animal shelter in the US, I will be investigating which characteristics have the highest correlation with successful adoptions. Further investigations will be performed to find additional correlations from the varying characteristics within the data. The primary data set that will be analyzed is a 8-year span data set consisting of approximately 125,000 data points with 11 features (City of Austin, Texas, 2021).

Using some of the most widely used machine learning models, a predictive outcome methodology would be followed to compute the likelihood of an animal being adopted. This project will serve as a tool for animal centers to determine if additional emphasis should be spent on those animals who are less fortunate and could use additional attention to help influence a successful adoption. With the interest to search for the major characteristics that influence the likelihood of an adoption for dogs and cats, additional analysis will be performed to find other correlations to factors that affect adoption rates. Having this additional knowledge why animals with certain characteristics have greater difficulty finding homes could be helpful for animal shelters to place additional effort and time when marketing them. For example, what are the most attractive names of pets that seem to be adopted? Would giving an animal a name such as “Pebbles” or “Cookie” entice lookers to research more about a prospective pet. The general steps that will be processed to develop create the analytic tool that will help forecast the likelihood of adoption will consist of two phases. The first phase will involve data exploration and data processing which involves cleaning of the data, removing incomplete entries, feature engineering, and assigning numerical values to the categorical data. The second phase will involve testing the data set with various machine learning algorithms such as random forest and logistic regression classifiers. After completing the second phase, further analysis will be made which involves comparison of false and true positive scoring before settling on a model. Having such tools developed will serve to benefit all types of animal shelters around the world and would not be exclusive to cats and dogs.

2.2 Significance of the Project

Having access to a powerful analysis tool powered by the latest machine learning models in combination with an ever-growing data set over time could be a valuable time, resource and a animal saving asset. Some of the obvious benefits include reduced animal shelter stays and less euthanasia’s performed. Having shorter shelter stays would allow for additional rescues to be made to other unfortunate animals instead of otherwise be turned away due to overcrowding. If a tool would grant benefits such as less euthanasia's being done in shelters, an additional benefit of reducing the burden that’s created on the staff who perform them. While the idea behind this project has been explored before, re-visiting this with the latest advancements could yield an improvement in accuracy or at the least verify confirmation of previous findings, due to the maturity of current machine learning models. This project would pave the way for further advancements that could involve what states or locations have poor adoption rates and determine what the tradeoff is between adoption time versus relocating the animal to another shelter with higher adoption success.

3.1 Previous Research

Previously there have been studies that looked at the correlation with the importance of names of the animals being kept in the animal shelter. The most common dog names were Max, Bella, and Daisy, while Ginger was the dog's name that resulted in the most adoptions. Among cat names, Cookie was the clear winner (Andrews, 2018). Many of the analysis's performed by previous data scientists used an older data set that consisted of approximately 27,000 data points, far fewer entries than what's available now to come up with findings. Using the findings previously discovered from the older data set will serve as a way to check if we're moving in the right direction with the newer analysis on a larger and more recently updated database.

There have also been studies investigating positive benefits of neutering pre-adopted animals based on the probability of animal adoption (Clevenger & Kass, 2003). The study discussed the impact of the free sterilization services offered by veterinary medical schools and how it increased the chances of pet adoptions. The collaboration between both parties yielded an overall lower euthanization of adoptable pets.

A similar study that was done related is the prediction of adoption versus euthanasia among dogs and cats in a California animal shelter (Lepper et al., 2002). The raw data set that was analyzed totaled ~17,000 records which eventually broke down to as little as 7720 dogs and 6011 cats that aided in the study. The final results of the study were that 26% of the dogs and 20% of the cats were adopted which was considerably lower than originally calculated. The takeaway from this study was there should be careful consideration for the availability of adoption before being considered an adoption candidate.

3.2 Findings and unanswered questions

Many of the studies conducted is based on relatively small data sets provided by handful of animal shelters around the United States. Having access to a considerably larger data set would be a way to solidify previous findings or the opposite. The current database by the Austin based organization now has over 125,000 data points which is approximately four times larger than what many early studies used. One of the limitations of this analysis is that the data that will be isolated to Austin Texas, with the positive that they're one of the few who have kept a tightly curated database of dog and cat adoptions.

3.3 Your preliminary work on the project

Before settling on this combination of machine learning and animal adoptability study, an extended amount of time was invested to check the feasibility of completing such a project in the required amount of time. Various studies and articles were collected to gain further understanding how the animal shelter industry worked. Having family friends who are active veterinarians who supported this study was another big reason for continuing on with the study. Having only been recently involved with the manipulation of data sets and tuning of machine learning models through recent graduate studies, this project will serve as the first exposure outside of those dedicated courses. While my experience with animal adoption centers is at a personal level, my inclination is pets that are older in age are among the highest reason's adoption would be difficult with the notion it's far easier to train a younger animal in a household. Next on the list would be single or black color coated animals on the rank to be least sought after.

3.4 Remaining questions

Attempting the use of one or more machine learning algorithms will be necessary to find new insights behind adoption rates including but not limited to: Which features affect the success of adoption? Which types of animals are sought after more so than others? What type of animals have the highest return or failure to keep after adoption? Which time of the day, month or year is the highest rate of adoption? While keeping the scope narrow, we'll first explore which factors lead to a successful adoption, before paying closer attention to what other variables within that feature attributes to poor adoption.

4.1 Approach

With the understanding that the use of Machine Learning as a tool is only as good as the quality of data that is fed to into it. This set my first goal to search for an available data set before considering any web scraping techniques. In my search for data, there were shelters with very detailed data sets that included as many as 23 features per animal such as how they were adopted, how long their stay was, and if it was a shelter transfer. Unfortunately, as detailed as some of the data sets were, these had as few as ~10,000 rows of entries. It wasn't until coming across a shelter based out in Austin Texas that consistently updates and releases their shelter data for the last eight consecutive years. This gave promise that there exists a reasonable amount of data to generate an accurate and predictable model for the project without too much difficulty. The animal shelter based in Austin provides shelter to ~18,000 animals pear year and is known as one of the largest "No Kill Shelter". The latest 2021 data set available has approximately 125,000 data points in the set. Having a large data set will be beneficial for achieving good model performance with the thousands of variables per class. Nonlinear classifiers such as random forest or logistic regression will benefit tremendously from this amount of data.

The first step to understanding the data is performing an exploratory data analysis to gain further understanding to the most obvious traits that impact adoption rates. Before taking a peek at the data, the entire set will be pre-processed and cleaned up removing any anomalies or major data points that could skew the analysis. Because we're dealing with a large data set, I would simply remove any of the anomalies such as missing labels and not attempt to artificially fill them. If there is a substantial number of rows with incomplete features, I will attempt to handle them accordingly with feature engineering techniques. The data will be looked at holistically before manipulating the data in segments. The next step will be to analyze each animal by type e.g., cats and dogs. Within each type there's multiple classifications for each animal that range from "Return to owner", "Euthanasia", "Adoption", "Transfer", and "Died". All Categorical features as well as the Classification will be one hot encoded while numerical features will be scaled accordingly. A feature that will be carefully processed will be the "age upon outcome", which currently has varying unit values in days, months, and years. This feature will be converted into only days with a limitation that converting one-year or two-year values will have a coarse increment of 365 days to 730 days or etc. after it exceeds a single year of measurement.

The data set will be split into a 70% Training, 20% Validation, and %10 Testing data set. How well the models perform in regards to classification accuracy, Precision, Recall and F1 score, additional post processing may be performed and cause for a re-visit to our "Feature Engineering" step. An example would be to add additional grouped features such as coat patterns of the pet. For example, for dog breeds of type dalmatian, a description of "white spotted" would be added. We could also create another grouped feature since the data lists the age of the animal by year. A new feature would be to indicate if the animal is a kitten/puppy, adult or even elderly.

Depending how well or poorly feature engineering is implemented, the newly created feature(s) may be thrown out to avoid adding complexity to the model and adversely affecting the analysis.

With the data set compiled, an evaluation of the size will determine which models would be used. The first machine learning technique will be using Random Forest Classifier as it’s known to perform exceptionally well with a data set as the one that’s being analyzed. Attempting to achieve the highest series of scores would be achieved by adjusting the various parameters such as “number of decision tree classifiers”, and “minimum amount of leaf nodes “. Conducting an analysis using a logistic regression classifier would also be performed and compared to the previous algorithm used. Attempting to yield the highest series of scores will be made through hyper parameter tuning steps. Both models will be used to test for accuracy against the test set. Depending on the results, the most accurate model would be used toward the final predictive tool in determining adoption success. Summarizing all the steps that was just discussed will follow a general KDD methodology as seen in Figure 1. The reasoning to staying consistent with a KDD methodology is at an abstract level, the KDD field’s primary focus is with the development of methods and techniques for making sense of data (Fayyad et al., 1996).

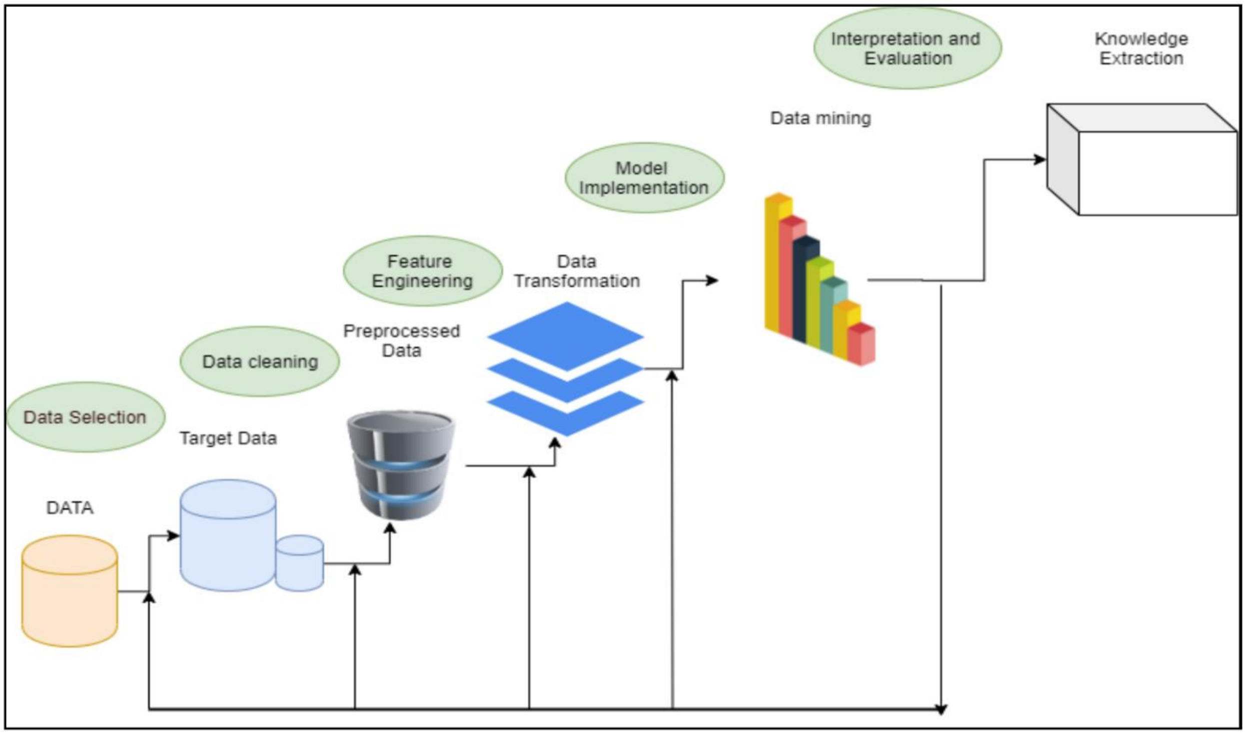


Figure 1: KDD Methodology

4.2 Data Collection

This section consists of the primary data set that was used for this capstone. At the time while working with the data set from Austin Animal Center Shelter, the total data set consisted of over 125 thousand inputs with a date range from 2013 to 2021. The input data was time stamped down the minute of the day. The reason for settling on the data set provided by the City of Austin was due to them being known for the largest no kill animal shelter with the largest publicly available data set. The data set was imported directly from the Austin animal web portal. Within the data set included features: animal id, name, datetime, monthyear, date of birth, outcome type, outcome subtype, animal type, sex upon outcome, age upon outcome, breed, and color. The feature that we will consider the target output variable will be outcome type which included: adoption, died, disposal, missing, relocate, return to owner, return to adoption, as well as unwanted labels that would need further processing to clean up. All other features outside of "outcome type" were used while training the classifier when building the machine learning model.

4.3 Analytical methods

Before building up any models the first step was to analyze the raw data points in the data set to see what we were working with. Upon searching for all the unique animal types entered into the database we found the data base consisted of cats, dogs, bird, livestock and other animals. To minimize any data confusion and keep the scope of the project minimal all animals other than cats or dogs will be removed from the data set. Removing all the unwanted animal types from the data set trimmed our total entries from 125 thousand down to 117 thousand shelter inputs. The breakdown between cats and dogs can be seen in figure 2.

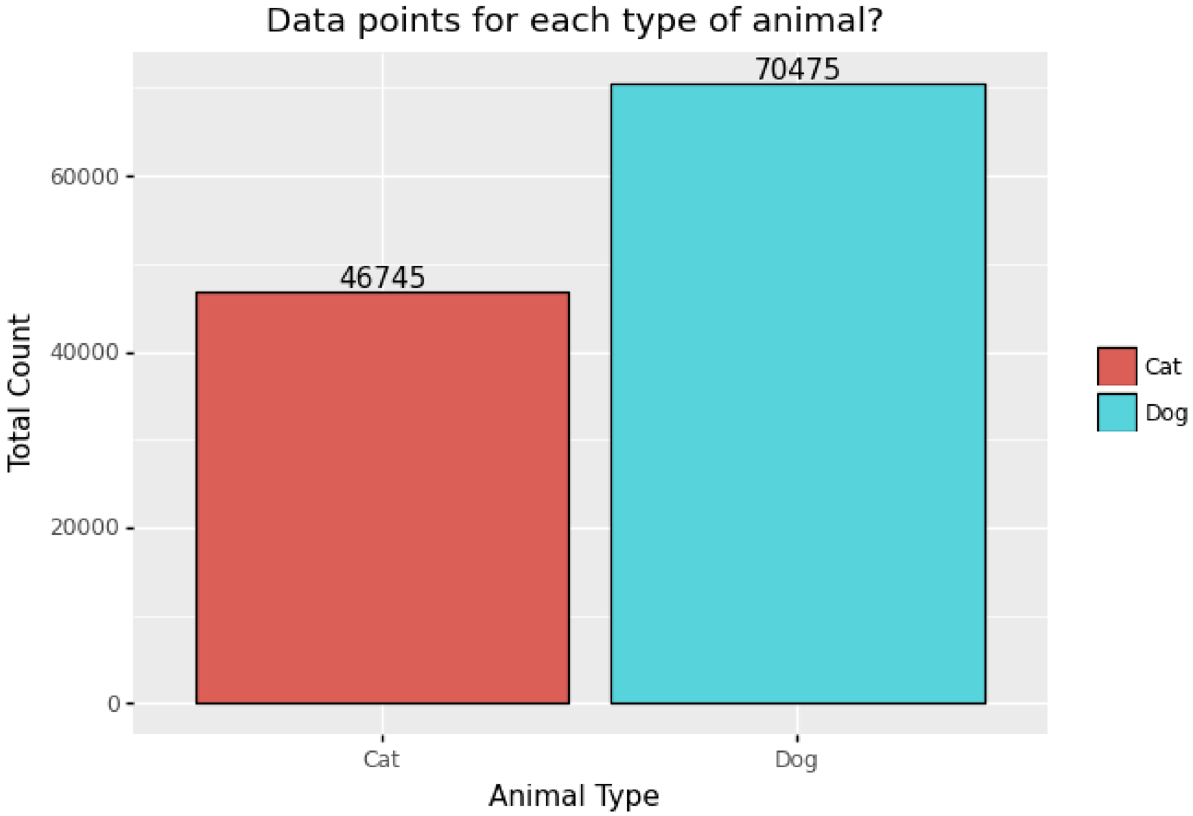


Figure 2: Total Count for Cats & Dogs

Further steps were then made to understand the outcome types broken down between cats and dogs. Figure 3 shows the breakdown between animals and gives us insight that cats seemed to be Adopted less, with less chances of returning to owners and have higher shelter transfers. While for dogs it was the opposite with higher chance of adoption, higher return rate to owners, and a lower ratio of transfers to that of cats.

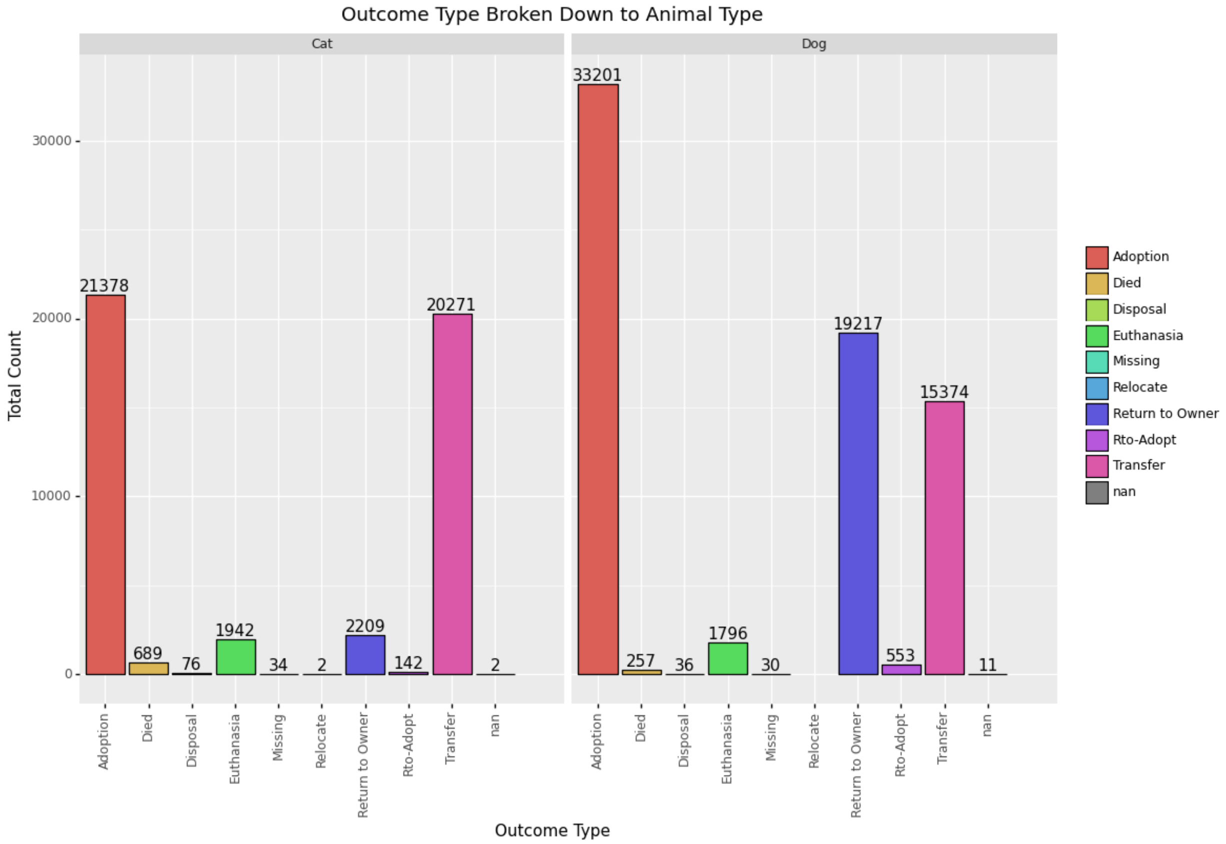


Figure 3: Outcome Type Broken Down Between Animal

Looking at figure 4, it's seen that one of the major features that have a correlation to adoption rates is whether or not the animal is spayed or neutered. Unfortunately, this may be an example where correlation does not imply causation. For reasons that further research found that a majority of states have implemented a mandatory spay and neuter law to help with the overpopulation of homeless animals (Hodges, 2010).

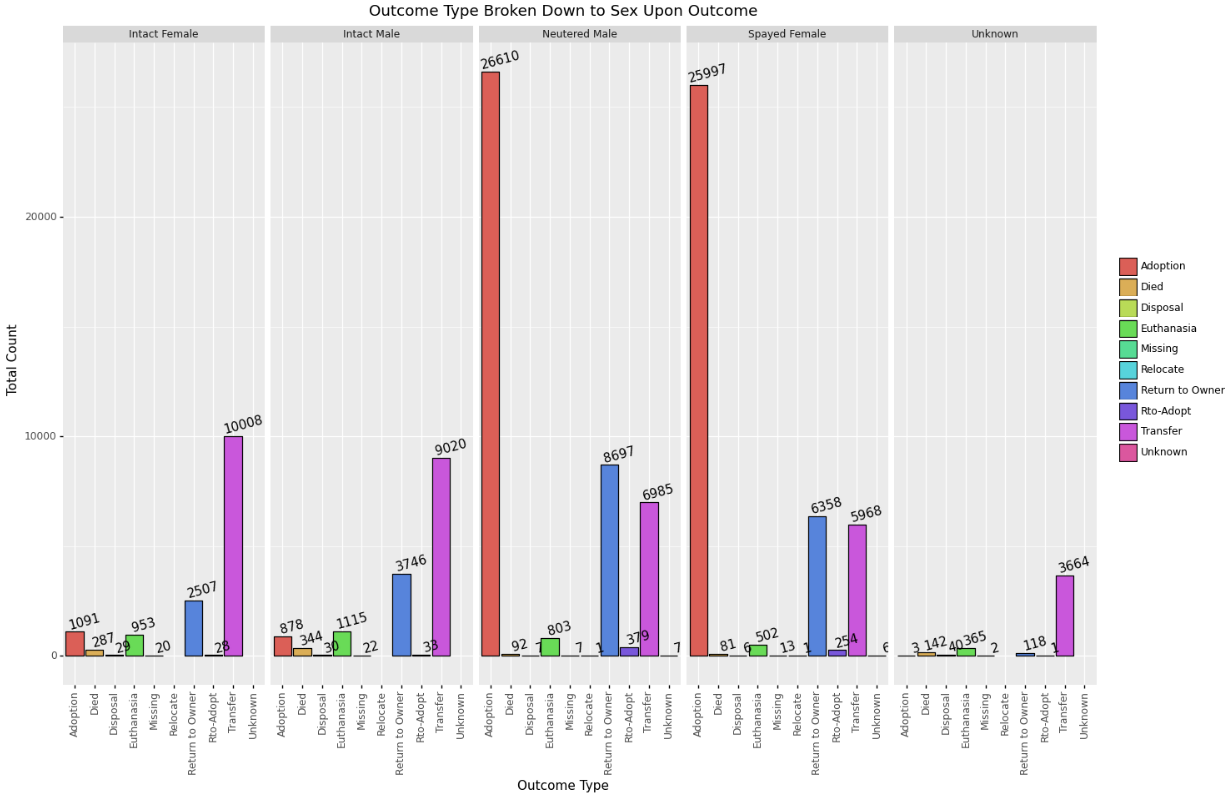


Figure 4: Outcome Type Broken Down to Sex Upon Outcome

Using the date time feature which is also the for time of intake for each data entry it was possible to perform a time domain analysis on the overall shelter activity shown in figure 5. From this we could see during the summer of every year activity slowly ramped up and spiked before settling down by the end of the year. This gives us further insight that shelters are very overwhelmed during these periods of the year and could use any bit of help to reduce the strain. An article pointed out that during the peak seasons of animal intakes is also the same time pet adoptions temporarily dip (Muñoz, 2019).

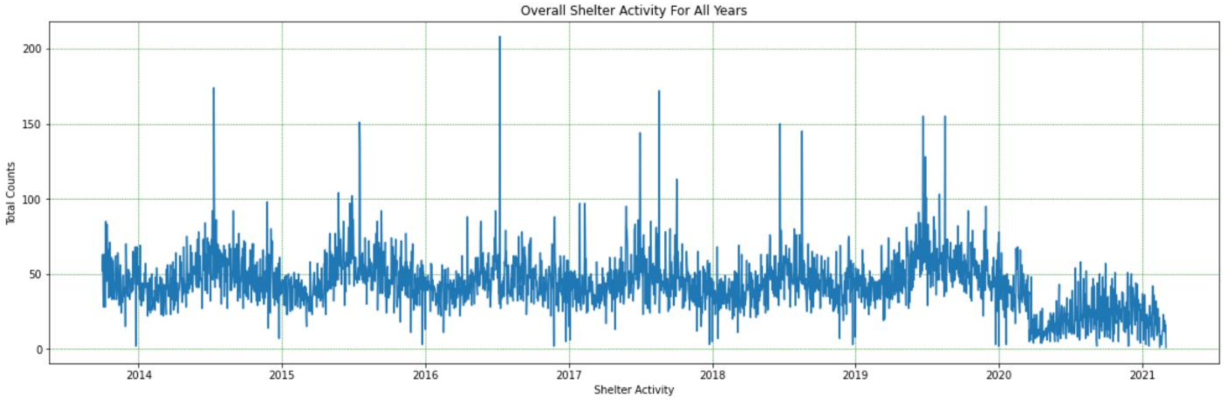


Figure 5: Overall Shelter Activity for All Years

Taking the overall shelter data and breaking it down into the possible outcomes: return to owner, return to adoption, transfer, adoption died, euthanasia we can chart the data over time seen on figure 6. To perform this step, all the data was filtered into a single day before it was charted. Judging from the data it seems that euthanasia’s performed, gradually lessened throughout the years while animal adoption stayed consistent. To simplify the time chart further to make it easier to see trends, further processing was made by showing the monthly activity instead of daily as shown in figure 7. Personally, I was unable to deduce any information from the graphs that could assist with improving the model.

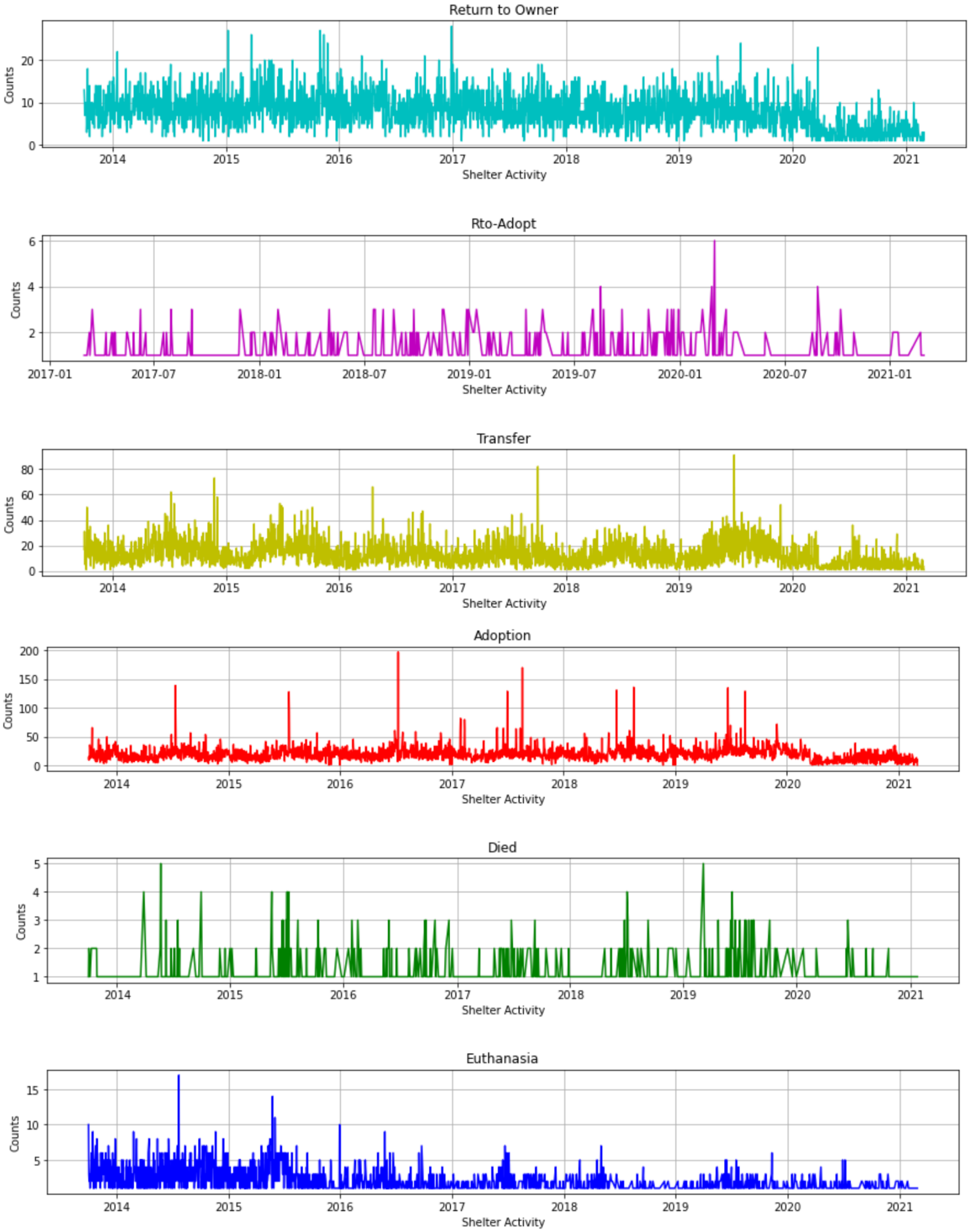


Figure 6: Shelter Data Broken Down to Outcome Types (Days)

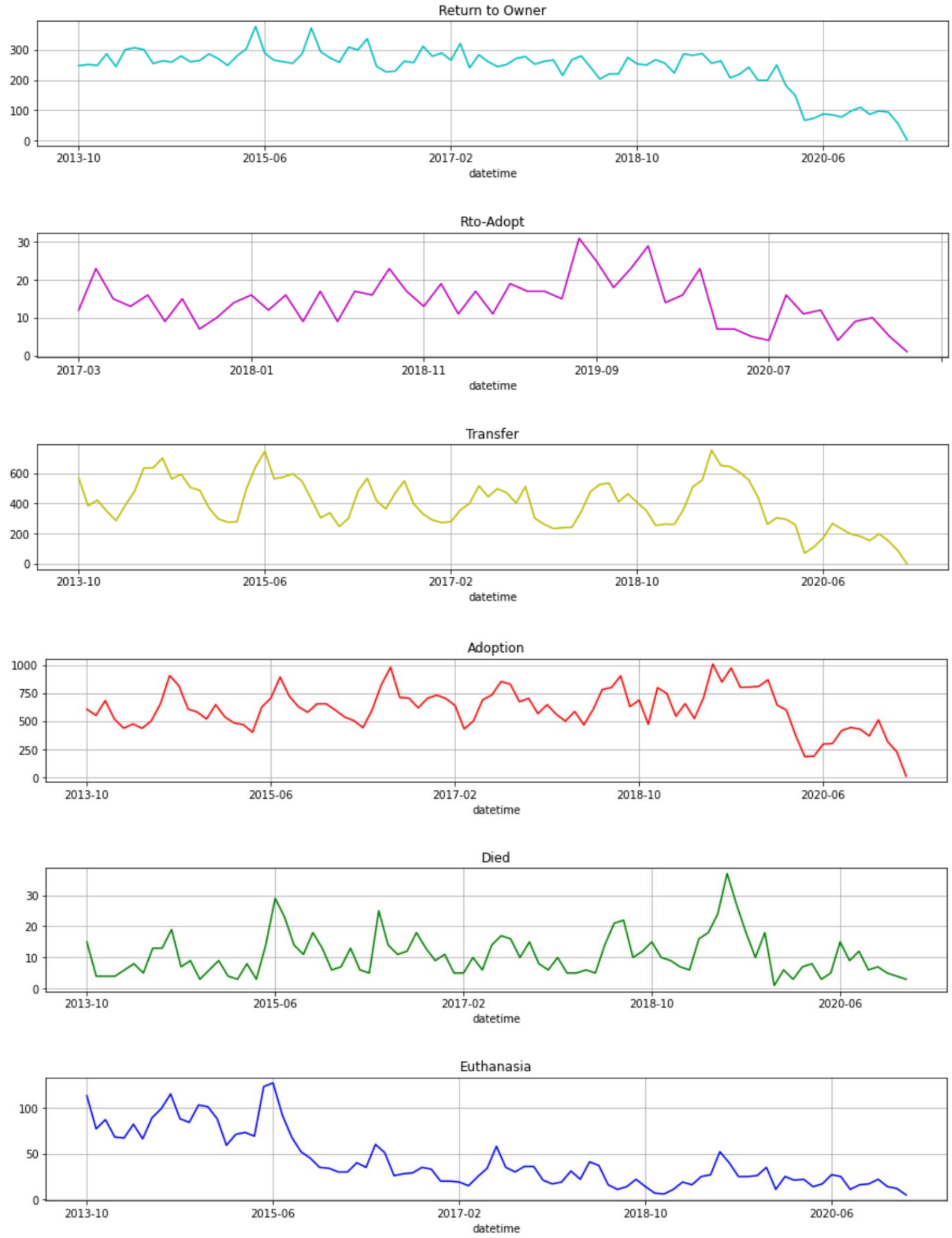
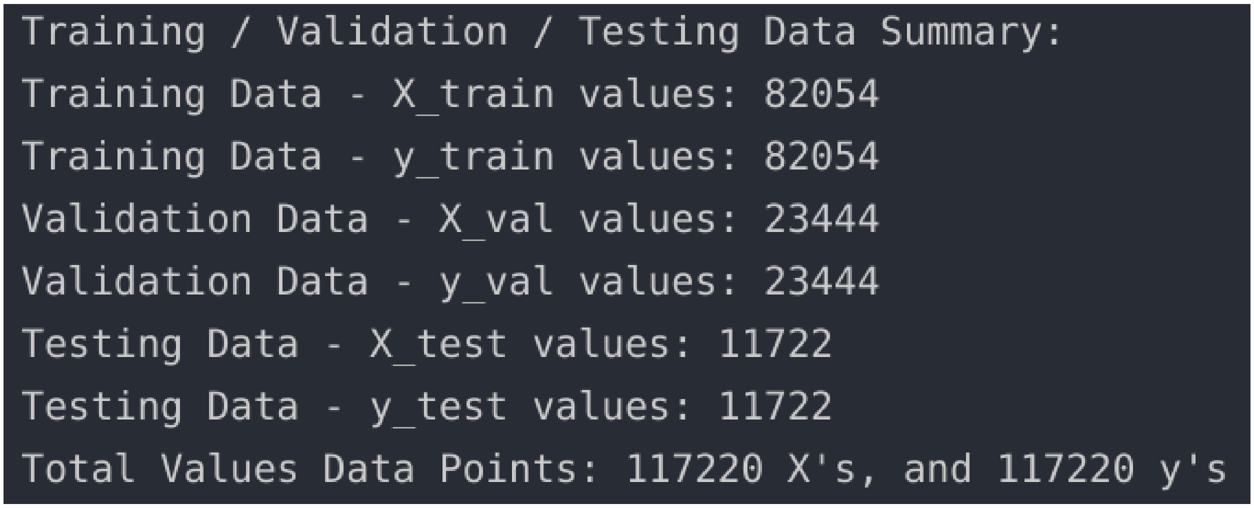


Figure 7: Shelter Data Broken Down to Outcome Types (Months)

After getting an idea what the data looks like, the next step was to perform a series of feature engineering enhancements. Some basic administrative steps were made to make it easier to work with the data such as replacing all spaces in the feature names with underscore and text to lowercase, dropped all animals that were not dogs or cats, and remove any entries that did not have any age outcomes. To keep the scope of the study relatively simple we added a new feature and set successful adoptions to a numerical “1” while anything other than adoption was a numerical “0”. Before moving further, the data was checked for any missing/empty data in a given row. The original date time for each entry was not suitable to be used for training due to its complex format of day-month-year-hour-minute-seconds-am/pm format. A simple function reduced this down to a calendar day for all data points. Another issue that needed to be resolved was the varying units used for age upon outcome. Instead of using units of years, months, week, or days, a function was created to convert all animal ages to unit days. Because we’re dealing with a large data set of categorical values, label encoding is used to convert them to numerical values in order for machine learning models to accept them as training input. To remove as much noise from the model, features such as animal id, index, and redundant features were removed. The adopted feature which was derived from the outcome type was transferred to the target output values to train the model against. At last, the data was split into a 70/20/10 split of Training, Validation, and Testing data. Figure 8 is the detailed breakdown after the data is processed through the split function.



```
Training / Validation / Testing Data Summary:
Training Data - X_train values: 82054
Training Data - y_train values: 82054
Validation Data - X_val values: 23444
Validation Data - y_val values: 23444
Testing Data - X_test values: 11722
Testing Data - y_test values: 11722
Total Values Data Points: 117220 X's, and 117220 y's
```

Figure 8: Training/Validation/Testing Data Breakdown

The first machine learning model that will be implemented is the random forest classifier. Random forest is an ensemble of learning algorithms ideally used for classification purposes through the use of multiple decision trees. After tuning the parameters of the model to achieve the highest perceived score, we settled with the estimators to 200 and max depth to 40. Estimators is the number of trees and max depth is the maximum depth or longest path between the root node and a leaf node. The second machine learning model that was used to compare to the random forest model is the logistic regression classifier. The tuning parameters that were settled on that model was allowing a max iteration of 1000 and the remaining parameters default to the library.

4.4 Plans for interpreting results

The results from the regression trees will serve as additional information to how much weight various features have on the likelihood towards adoption. One of the difficulties that will be encountered is interpreting the series of decision trees that was generated for the random forest classifier due to the maximum depths of 40 and 200 estimators set. While previous studies performed basic decision/regression trees it was less trivial to chart and analyze the model's process flow. The results from this analysis would give insight to which areas to focus on within the data set and possibly improve the model. For example, if there seems to be substantial value or lack of a given feature, re-visiting the data set and implementing additional improvements to the feature engineering steps could be made.

4.5 Results

To gauge the performance of the model we scored for accuracy, precision, recall and f1 score each time the model was trained/fitted. Accuracy: is how many of the predictions were correct, Precision: out of all the truth predictions, how many were correct, and recall: is out of all the possible truths, how many were correct. Closer attention is given to the F1-score which is a combination of both precision and recall to ultimately measure the accuracy of the model (Riggio, 2019). The results from the random forest classifier are shown in figures 9 and 10. True negative of 5628 versus it's False negative of 533 gave us an accuracy score of 91%. While True positive was 4895 with a false positive of 666 yielded an accuracy of 88%. Precision, Recall and F1-Score averaged out to be approximately 90% which is surprisingly accurate in predicting the success of adoption.

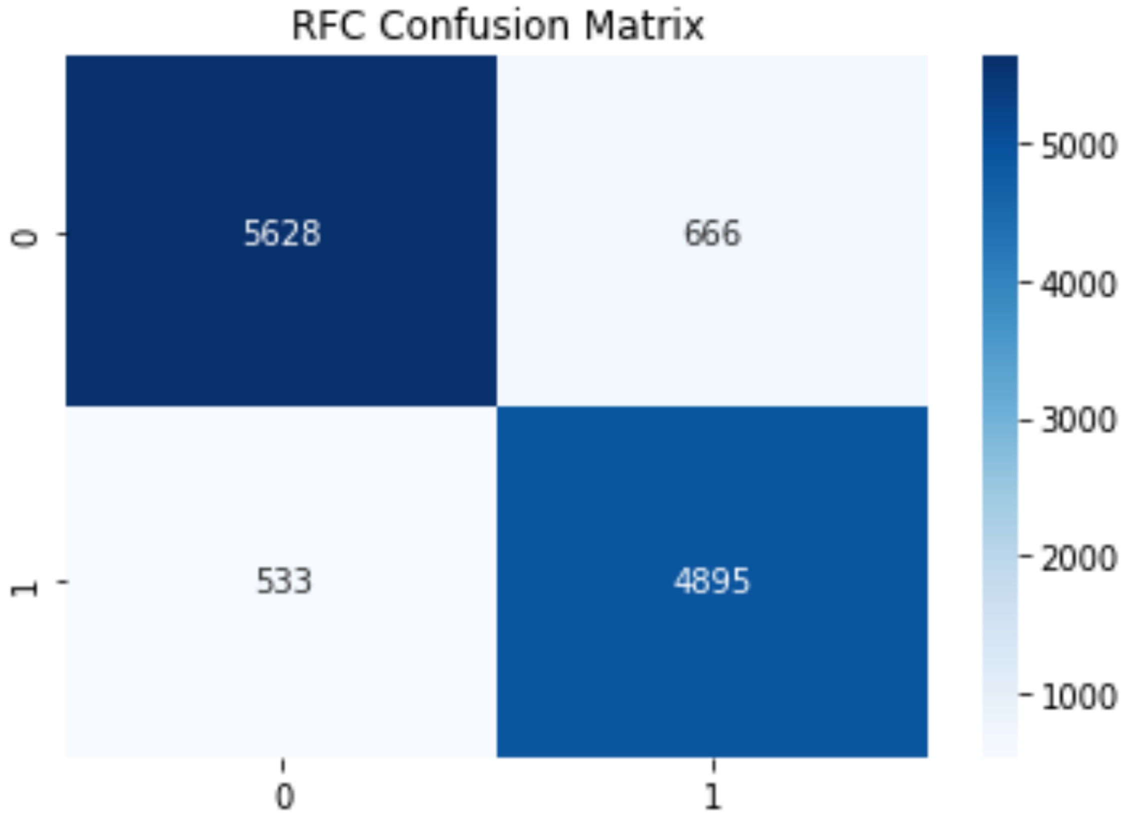
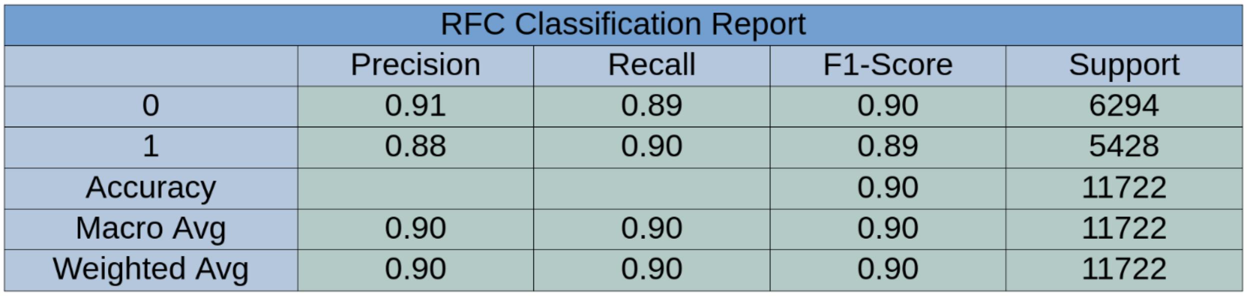


Figure 9: Random Forest Confusion Matrix



RFC Classification Report				
	Precision	Recall	F1-Score	Support
0	0.91	0.89	0.90	6294
1	0.88	0.90	0.89	5428
Accuracy			0.90	11722
Macro Avg	0.90	0.90	0.90	11722
Weighted Avg	0.90	0.90	0.90	11722

Figure 10: Random Forest Classification Report

The results from the logistic regression classifier are shown in figures 11 and 12. True negative of 4229 versus it’s False negative of 1697 gave us an accuracy score of 71%. While True positive was 3731 with a false positive of 2065 yielded an accuracy of 64%. Precision, Recall and F1-Score averaged out to be approximately 68% which is substantially lower than what we scored for the random forest classifier.

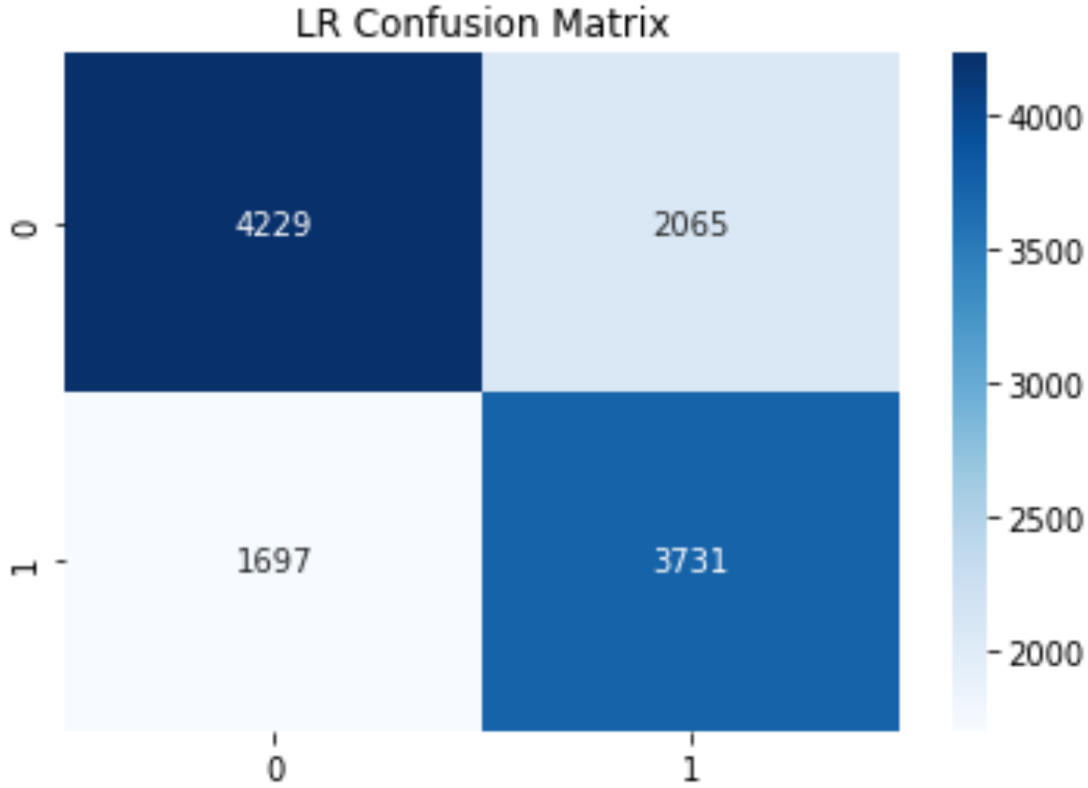
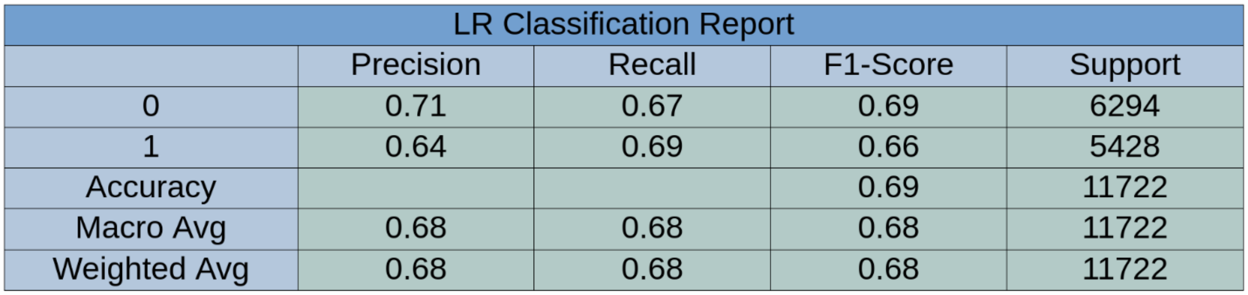


Figure 11: Logistic Regression Confusion Matrix



LR Classification Report				
	Precision	Recall	F1-Score	Support
0	0.71	0.67	0.69	6294
1	0.64	0.69	0.66	5428
Accuracy			0.69	11722
Macro Avg	0.68	0.68	0.68	11722
Weighted Avg	0.68	0.68	0.68	11722

Figure 12: Logistic Regression Classification Report

Comparing the results from the two classification algorithms it was determined to settle with the random forest classification. Figure 13 visually displays the associated weights for the various features that the random forest model perceives. The most important predictor in predicting the outcome seemed to be the outcome subtype which consists of variables such as behavior, in foster, aggressive, suffering, and etc. Second to outcome subtype was sex upon outcome and next was age of the animal which is consistent with the idea that younger animals tend to be adopted more easily. This is the case due to senior pets, even those that are a few years old normally languish in shelters because people prefer a younger pet (Brown, 2020).

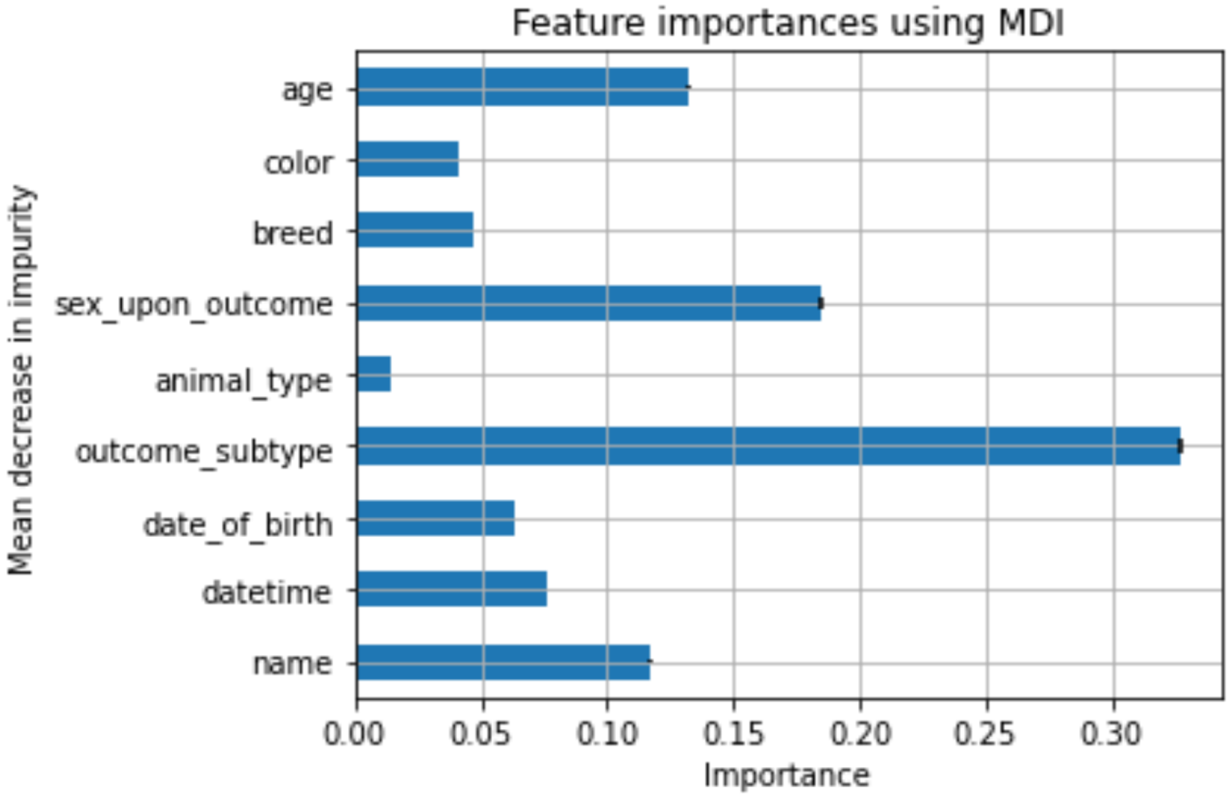


Figure 13: Random Forest Feature Importance's

6.1 Discussion

Our primary research goal was to find a suitable model that could accurately predict the outcome of animals to determine if additional assistance is necessary to help those that are not as fortunate. The complete analysis also was able to find correlations to what factors have the highest weight in determining adoption likelihood. While figure 13 gives sex upon outcome a relatively high weight, it may not be as important due to the inner workings and operations of how shelters following local spaying and neutering laws. Looking back at the experimentation

]

process, having reduced the estimators and depth to a minimal single digit value would have presented a clearer graphic representation of how the model is behaving. While the full random forest model correctly identifies ~90% of the animal adoptions, further work should be done to verify the validity of the accuracy before accepting it at face value. The logistic regression model’s ~68% score was in line with what other researchers came across when using similar models and data sets.

6.2 Conclusion

The results that we’ve concluded to can be used by the shelter to help narrow down which animals would have difficulty being adopted and make changes to the resources spent. With this initial study set, our model could be used as a baseline to look into other variables such as how long an animal is kept in shelter, intake age, and even outcome age. While the high score that the random forest classifier achieved was very promising, it’s not clear where the decision thresholds are without extracting all the trees generated by the algorithm and analyzing them together. An attempt was made to reduce the total estimators and maximum depth, but performance of the model drastically suffered to where it performed on par or worse than the logistic regression model. Another improvement that could be made to the existing model would be to creating filter categories for some of the features that have large amounts of unique variables such as color or breed of an animal. While the color featured displayed over 500 unique values, the breed feature had over 2,000 varying breeds. Creating a comprehensive filter function would have taken more time allotted but may serve to help improve the consistency of the model.

References

Andrews, A. (2018, October 31). *How Data Science Can Be Used to Improve Animal Shelter Outcomes*. https://www.switchup.org/blog/how-data-science-can-be-used-to-improve-animal-shelter-outcomes

ASPCA. (2017, March 10). *ASPCA Releases New Data Showing Remarkable Progress for Homeless Dogs & Cats*. ASPCA. https://www.aspca.org/about-us/press-releases/aspca-releases-new-data-showing-remarkable-progress-homeless-dogs-cats

Bradley, J., & Rajendran, S. (2021). Increasing adoption rates at animal shelters: A two-phase approach to predict length of stay and optimal shelter allocation. *BMC Veterinary Research*, *17*(1), 70. https://doi.org/10.1186/s12917-020-02728-2

City of Austin, Texas. (2021). *Austin Animal Center Outcomes: Austin Animal Center Outcomes* [Application/xml]. City of Austin, Texas Open Data. https://data.austintexas.gov/d/9t4d-g238

Clevenger, J., & Kass, P. H. (2003). Determinants of adoption and euthanasia of shelter dogs spayed or neutered in the university of california veterinary student surgery program compared to other shelter dogs. *Journal of Veterinary Medical Education*, *30*(4), 372–378. https://doi.org/10.3138/jvme.30.4.372

Evenblij, K., Pasman, H. R. W., van Delden, J. J. M., van der Heide, A., van de Vathorst, S., Willems, D. L., & Onwuteaka-Philipsen, B. D. (2019). Physicians' experiences with euthanasia: A cross-sectional survey amongst a random sample of Dutch physicians to explore their concerns, feelings and pressure. *BMC Family Practice, 20*(1), 177. https://doi.org/10.1186/s12875-019-1067-8

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–37. https://doi.org/10.1609/aimag.v17i3.1230

Hodges, C. (2010). *State Spay and Neuter Laws | Animal Legal & Historical Center*. https://www.animallaw.info/intro/state-spay-and-neuter-laws

Lepper, M., Kass, P. H., & Hart, L. A. (2002). Prediction of adoption versus euthanasia among dogs and cats in a California animal shelter. *Journal of Applied Animal Welfare Science: JAAWS*, *5*(1), 29–42. https://doi.org/10.1207/S15327604JAWS0501_3

Muñoz, A. (2019, May 15). *Why the animal shelter population soars during summer months*. Miami’s Community News. https://communitynewspapers.com/doraltribune/why-the-animal-shelter-population-soars-during-summer-months/

Riggio, C. (2019, November 3). *What’s the deal with Accuracy, Precision, Recall and F1?* Medium. https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021

Sencer. (2017, May 24). *Do No-Kill Shelters Really Benefit Animals?* KQED. https://www.kqed.org/education/499450/do-no-kill-shelters-really-benefit-animals

Appendix

Python Notebook Code Location: https://github.com/jsand153/ENG_296.git