**Data Mining Final Project Report: Analysis of Pet Adoption Data for Prediction Modeling**

Sheree Harris

E-Mail: slh141@students.uwf.edu

**Abstract**

In the last four or so years, pet shelters nationwide have been feeling the strain of overcrowding (Bautista-Alejandre, 2024). While the overall amount of pets or strays being taken into the shelters has generally decreased, so too has the number of adoptions resulting in an increase in the length an animal stays at the shelter (Teich, 2024). Therefore, new solutions need to be found in order try to alleviate this burden on the shelters.

1. **Introduction**

In order to address this issue, I chose to focus on the data from the pet shelters themselves. Specifically, I looked into the various factors that a pet shelter is likely to record and looked to create a program to find the trends within. My hope was to figure out what characteristics in a pet might a potential adopter look at, and what factors might increase the chances of pets getting adopted. Then from this data the program could more accurately predict what pets were more likely to get adopted.

Given that there are thousands of pet shelters in the U.S., I wanted to get a large sample of data to use with out having to worry about formatting difference and gaps in the data. I also wanted to make sure the data covered a decent period of time and was from a reputable source. After looking through Kaggle and data.gov, I found the data for the Austin Animal Center. This shelter is “no kill” and thought to be one of the largest shelters of its kind in the country (Austin Animal Center, n.d.). The data came in two parts, an intake dataset and an outcome dataset. Both datasets contained several factors concerning the pet at time of intake or when the pet left, with the pet sharing an animal ID across both of the datasets. This made this particular dataset useful when looking at it with my project in mind. A limitation of this, however, was that the data came from a singular source, so if other shelters collected data differently the results might not be as valid or helpful to them. Overall, however, I felt that due to the factors the dataset contained were fairly standard, the impact of the single source was minimal. One other limitation was that the dataset did not have a factor about size so, I could not see whether the size of the animal was a factor. This was slightly offset by the breed factor, but did not account for overweight animals.

By creating a prediction model based off of this data, there is the potential for it to help drive solutions to the issue of overcrowding in animal shelters. These solutions could help to increase adoption rates and shorten the length of time an animal is at the shelter, thus decreasing the strain on the shelter. With a decrease in strain, the shelters would be able to turn their attention to the animals who have been there the longest and provide them with the care they need, potentially getting them to their forever home after a far-to-long wait.

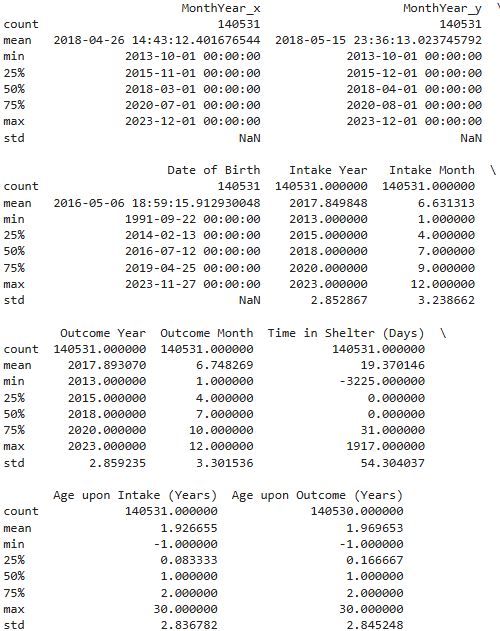
1. **Data Collection and Preprocessing**

As mentioned in the previous section, the Austin Animal Center data came in the form of two datasets, intake and outcome. Each dataset contained 157,628 rows of data and animal ID and name columns. The intake data included columns for the date when the anima arrived, the animal’s condition, breed, color, sex, and age as well as a few others. The outcome dataset had similar columns in breed, color, sex upon leaving (showing if the animals neutered status had changed when leaving), but also included the outcome of the animal and the date of when they left.

To begin handling the data, I first uploaded it into a MySQL database and then pulled the data into a Jupyter notebook for preprocessing. My main focus when preprocessing the data was to handle it as one dataset as opposed to the two it come from and then clean it up. So, I started by joining the dataset by the animal ID columns. Given that both datasets had some of the same columns I made sure that the joined dataframe handled them by adding an x or a y to the end of the name to keep the data separate. I then formatted the dates so that I could make a new feature, Time in Shelter (Days). This was so I could easily look at how long an animal had been at the shelter and see if there were any contributing factors to the length of the stay. Following that, I began to set the data up to be filtered. I dropped rows with missing data in what I believed were critical features, like animal type, breed, and the newly created Time in Shelter. Since some of the animals had been in and out of the shelter multiple times, I filtered it to only give the latest intake/outcome of the animal. With my data cleaned, I moved onto analyzing it.

1. **Data Analysis**

When analyzing the data, I decided to first get the statistics of the data to see if it reveal anything useful. This didn’t provide anything useful, so I decided to look at specifics, namely the outcome types and the types. From this I was able to see how many animals were actual adopted and gain some insight into what animal was the largest population.

A screenshot of a computer

Description automatically generated

Figure 1 Summary Statistics Figure 2 Specific Statistics

I then looked into seeing what trends I could find in the data. My main goal was to be able to visualize what the trends where to be able to better analyze them. I also wanted to highlight what trends existed in the features that were most likely to be shared amongst different shelters. For this I visualized the distribution of outcome types and then looked at animal types and time in shelter filtering for those that were actually adopted.

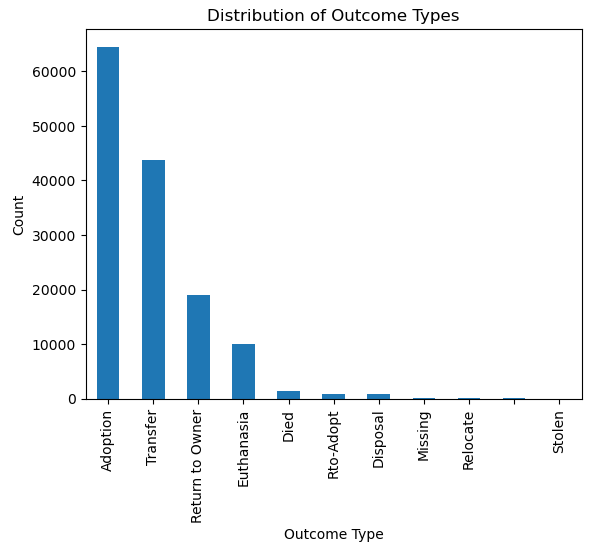


Figure 3 Outcome types distribution

A graph of adoption

Description automatically generated

Figure 4 Animal type by adoption

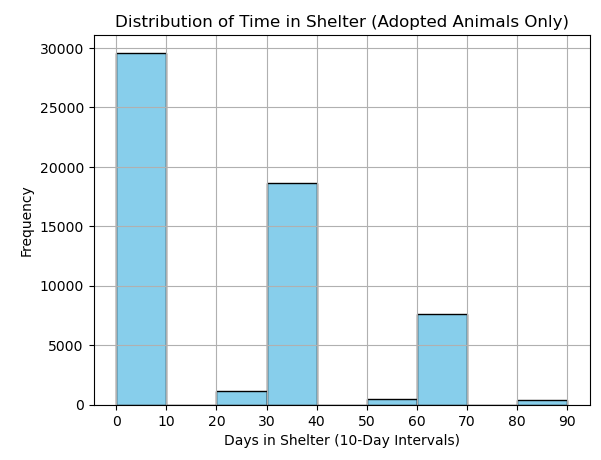


Figure 5 Time in Shelter by adoption

This showed me, first and foremost, that more animals than not were being adopted. It also showed me that more dogs were being adopted over cats, and that both were significantly more likely to be adopted than the other types of animals. However, part of that was likely due to the fact that the number of dogs and cats at the shelter was vastly greater than those of the other kinds of pets. The final trend of Time in Shelter by adoption showed me that most animals were adopted quickly, but the longer the animal was there the less likely they were to be adopted.

I then looked at the data to see what factors might be influencing the trends, looking at those features that would exist across all, or most, shelters. These features were breed, coloring, animal age, and the animal’s sex (i.e. neutered or intact).

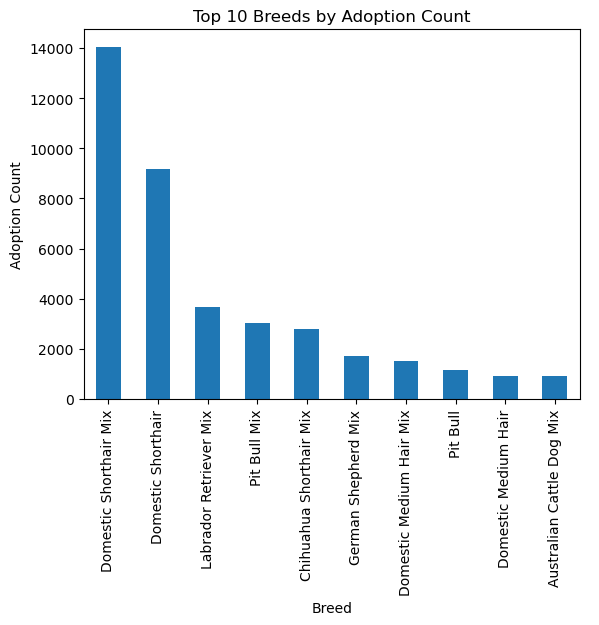


Figure 6 Top 10 Adopted Breeds

A graph of different colors

Description automatically generated

Figure 7 Top 10 Colors adopted

A graph of adoptions by sex

Description automatically generated

Figure 8 Sex of Adopted Animal

Looking at the displayed data I noted that despite more dogs being adopted than cats, the most popular breed was in fact the domestic short hair cat. I could also see that animals with darker colors were more likely to be adopted. These two insights could point to adopters preferring animals that either don’t shed as much, or whose shedding is less noticeable. Then looking at the animals’ neutered or intact status, it was clear that most animals adopted were neutered. One thing I didn’t take into account that might have provided valuable insight was whether the animal was already neutered upon intake or was neutered afterwards. If most of the animals came to the center already neutered, then most of the animals being neutered at adoption isn’t all that insight. Whereas, if they were neutered upon adoption would show that people don’t prefer intact pets.

One last category I wanted to analyze was time to see if there were any correlations there. For this I started big and then went smaller and more specific. I first looked at the trends over the years before narrowing it down to see if there were any months where more adoptions occurred than others.

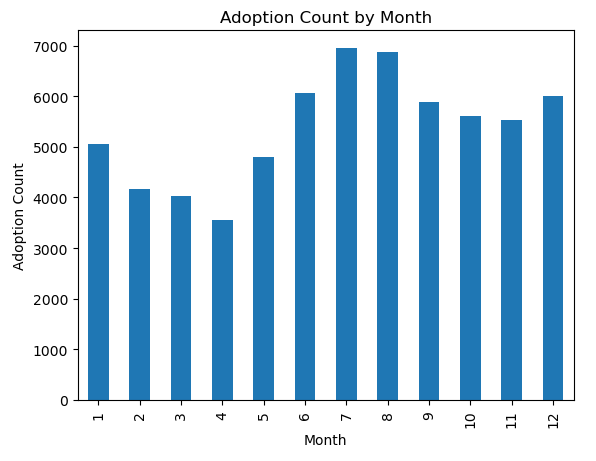


Figure 9 Months with most adoptions

The summer months saw the most adoptions, along with slightly higher adoption rates in the last half of the year. This could be attributed to people generally having less money at the beginning of the year due to the holidays and then having money again after the tax season ended. Surprisingly, the number of adoptions in December were not as high as I initially thought they would be, as I figured more parents would choose that time to get their children the pets they had been begging for all year.

The final time-based analysis I looked at was to see what correlation, if any, could be made between the age of the animal and the time they spent in the shelter. The correlation matrix revealed that there was no correlation between the age of the animal and time in the shelter. The time in shelter correlated with intakes and outcomes yielded low positives for both, showing that age wasn’t a factor. The perfect positive correlation between intake age and outcome age gives further credence to age not being a factor as this shows that the animals do not have a significant age difference from when they arrived.

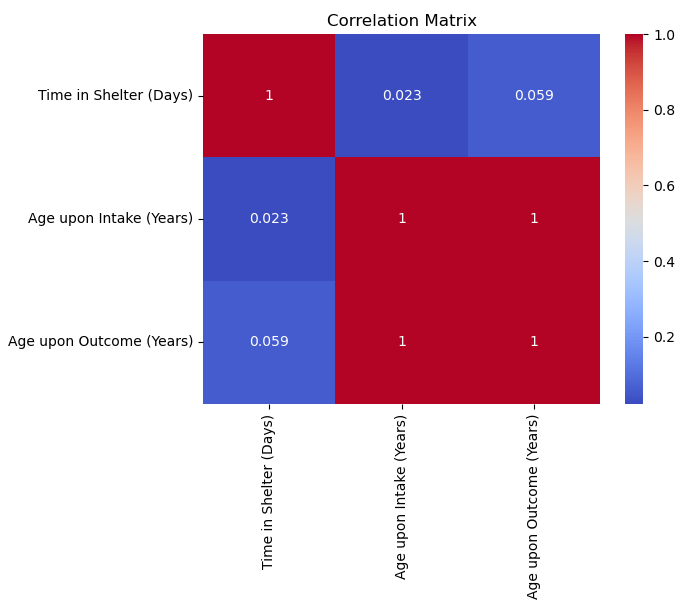


Figure 10 Time in Shelter vs Age

1. **Prediction Modeling**

With the data analyzed and some insights gained, I moved on to training and testing prediction models to see what method worked best. But first, I prepared the data to be run through the prediction models I chose, logistic regression, decision tree classifier, and random forest classifier.

To prepare the data, I made sure that it looked at only the records of those adopted since that was what I wanted to predict. I also made sure to target those features I had already “targeted” as being most likely to be reflected in the data from other shelters. The data was then split into sets for training and testing. After the data was prepared, I then ran it through each prediction model and then did cross-validation for each model.

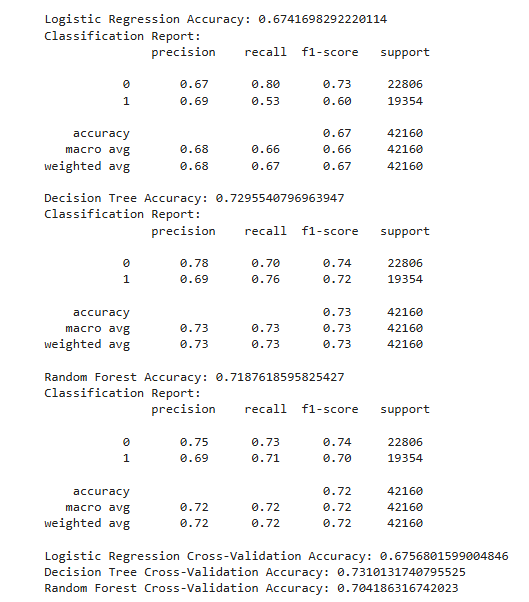


Figure 11 Prediction Modeling and cross-validation

The decision tree classifier was the most accurate of the three models, both before and after cross-validation, with a 73% accuracy. This was a little lower than I had hoped for, I would have preferred something in the 80% range. However, 73% is still fairly accurate and it was only tested on a single source of data. Since the decision tree classifier was the most accurate model, I used it to find the most important features when it came to predicting whether an animal would be adopted or not.

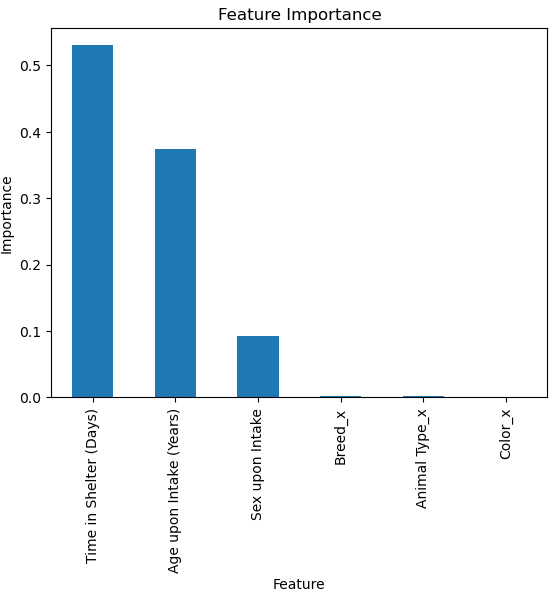


Figure 12 Important features

1. **Insights After Analysis & Conclusion**

Using the features deemed most important by prediction modeling, I went back and looked at the data analysis and the insights gained in order to compare what had been gleaned from the analysis. I found that the features I originally thought would have the biggest impact didn’t actually have that much sway according to the prediction modeling. Instead, those features were more inclined to show the adopters preference for what they were going to adopt rather than if they would adopt or not.

The time factors were the features that had more weight behind them when it came to whether adoption would occur. This tied back into the previous insight gained when looking at the time-based trends, that animals that had been there for a shorter time were more likely to be adopted than those who had already been there for a while.

With these insights in mind, to get more animals adopted shelters could use the preferences found in the data for marketing purposes. They could use those types of animals on the flyers and advertisements to draw a larger crowd to the shelter, more people coming to the shelter increases the chances of an animal getting adopted. Using the time-based trends, marketing could ramp up at the beginning of the summer months attract people during the months already shown to have a higher draw. The prediction modeling could then be used to show how effective the shelter’s campaign had been and whether they need to refocus elsewhere. Once the stain overcrowding has been relieved, the shelters can begin targeted campaigns to focus on the animals least likely to be adopted towards the end of the summer months going into the holiday season. This strategy would be done cyclically every adoption “season”. If the efficiency goes down, the program could be redone to show how the adopter’s preferences have changed and if the most important features remain the same.

In conclusion, I discovered that while adopters might have a preference when it comes to what they adopt. If they intend to adopt, they will adopt, generally going for the animals that haven’t been at the shelter for a long time regardless of the animal’s age. More research would need to be done to determine why the time in the shelter matters as much as it does. This could then be incorporated into the project to further enhance accuracy and available insights through data analysis.

# References

Austin Animal Center. (n.d.). *Austin Animal Center Data.* Retrieved from kaggle: https://www.kaggle.com/datasets/thedevastator/austin-animal-center-data?resource=download&select=Austin+Animal+Center+Intakes

Bautista-Alejandre, A. (2024, July 29). *New report reveals shelter pet adoption data.* Retrieved from dvm360: https://www.dvm360.com/view/new-report-reveals-shelter-pet-adoption-data

Teich, D. (2024, August 29). *Animal shelters in 2024: Challenges, opportunities, and tails of success.* Retrieved from ManyPets: https://manypets.com/us/blog/animal-shelters/

[Figure 1 Summary Statistics Figure 2 Specific Statistics 3](#_Toc184065401)

[Figure 3 Outcome types distribution 4](#_Toc184065402)

[Figure 4 Animal type by adoption 5](#_Toc184065403)

[Figure 5 Time in Shelter by adoption 5](#_Toc184065404)

[Figure 6 Top 10 Adopted Breeds 6](#_Toc184065405)

[Figure 7 Top 10 Colors adopted 7](#_Toc184065406)

[Figure 8 Sex of Adopted Animal 7](#_Toc184065407)

[Figure 9 Months with most adoptions 8](#_Toc184065408)

[Figure 10 Time in Shelter vs Age 9](#_Toc184065409)

[Figure 11 Prediction Modeling and cross-validation 10](#_Toc184065410)

[Figure 12 Important features 11](#_Toc184065411)