Prediction and Analysis of Animal Shelter Intakes and Outcomes

- Srimedha Bhavani Chandoo

- STAT 5600

Abstract

This report delves into the analysis and prediction of animal shelter intakes and outcomes using data from the Austin Animal Center. The dataset provides valuable insights into the shelter's operations, including the demographics of animals, the reasons for their intake, and the final outcomes such as adoption, transfer, or euthanasia. By employing exploratory data analysis (EDA) and statistical modeling, the report aims to identify patterns, optimize shelter operations, and enhance animal welfare.

Various models and statistical methods are utilized, including Maximum Likelihood Estimation (MLE), Ordinary Least Squares (OLS), Logistic Regression, K-Nearest Neighbors (KNN), AdaBoost, Gradient Boosting, and XGBoost. Based on accuracy scores, XGBoost delivers the best performance, followed closely by Gradient Boosting. These models provide a robust foundation for understanding outcome predictors and improving decision-making processes at the shelter. Key findings highlight factors influencing adoption likelihood, time spent in the shelter, and the impact of specific programs like Shelter-Neuter-Release (SNR).

Introduction

Animal shelters play a critical role in ensuring the welfare of vulnerable animals, providing them with care, medical attention, and opportunities for adoption. Among these institutions, the Austin Animal Center (AAC) stands out as the largest no-kill animal shelter in the United States. Each year, it provides care and shelter to more than 18,000 animals, embodying a commitment to compassion and ethical practices. The Center's mission is not just about housing animals but also about finding them suitable homes, offering rehabilitation, and engaging with the community to promote responsible pet ownership.

The dataset analyzed in this report is part of the Austin Animal Center's effort to share its operational insights through the city of Austin's Open Data Initiative. This dataset combines records of animal intakes and outcomes since October 2013, offering a wealth of information for understanding shelter operations. It includes key details such as animal species, breeds, ages, intake conditions, and final outcomes. The dataset provides an opportunity to explore questions like: What factors influence adoption rates? Which animals are more likely to face unfavorable outcomes? And how do seasonal trends affect shelter intakes and outcomes?

The importance of analyzing such a dataset extends beyond academic curiosity. Shelters often operate under resource constraints, needing to balance their no-kill policies with the realities of limited space, funding, and staffing. By leveraging data, shelters like AAC can make informed decisions to optimize their operations. Insights from this analysis can:

- Improve adoption rates by identifying characteristics that make animals more adoptable.
- Reduce the time animals spend in shelters by predicting and addressing bottlenecks.
- Support programs like Shelter-Neuter-Release (SNR), which aim to manage populations humanely.

Furthermore, understanding the patterns in intakes and outcomes can inform public policies, drive community engagement, and foster collaborations with partner organizations.

This study focuses on three main objectives:

1. **Intake Analysis**:
   - Examining the types of animals entering the shelter and their reasons for intake (e.g., strays, owner surrenders).
   - Understanding seasonal and demographic patterns in intakes.
2. **Outcome Analysis**:
   - Investigating the outcomes for animals, such as adoption, transfer, euthanasia, or Shelter-Neuter-Release.
   - Identifying factors influencing different outcomes, including animal type, age, and health conditions.
3. **Modeling and Prediction**:
   - Using statistical and machine learning models to predict outcomes based on intake data.
   - Comparing the performance of models like Logistic Regression, KNN, Gradient Boosting, and XGBoost to identify the most reliable predictors.

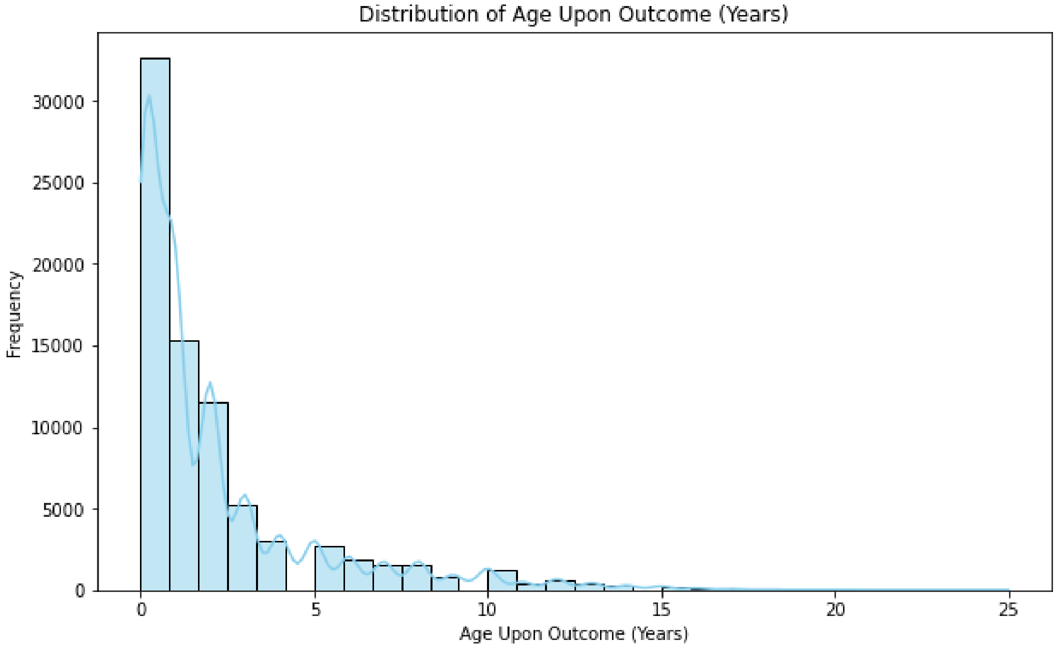
Animal shelter datasets often come with their own set of challenges. The AAC dataset is no exception, as it includes duplicate entries for animals that re-enter the shelter multiple times. This duplication, while meaningful, requires careful handling to avoid skewed analysis. Additionally, missing data in key attributes such as outcomes necessitates imputation or exclusion, which can impact the robustness of the results. Despite these challenges, the richness of the dataset offers immense potential for deriving actionable insights.

This study uses a mix of statistical techniques and advanced machine learning models to better understand and predict outcomes for animals in the shelter. Traditional methods like Maximum Likelihood Estimation (MLE) and Ordinary Least Squares (OLS) provide a solid foundation for identifying relationships between variables, while machine learning models such as Logistic Regression, K-Nearest Neighbors (KNN), AdaBoost, Gradient Boosting, and XGBoost offer more refined predictions. Among these, XGBoost has proven to be the most accurate, closely followed by Gradient Boosting, making them particularly effective tools for understanding and predicting outcomes. These models go beyond numbers—they offer practical ways to make better decisions. For example, they can help shelter staff focus on animals with the highest chances of adoption or identify those who might need extra attention due to longer stays. By providing actionable insights, these approaches ultimately support the shelter's mission to improve the lives of the animals in their care.

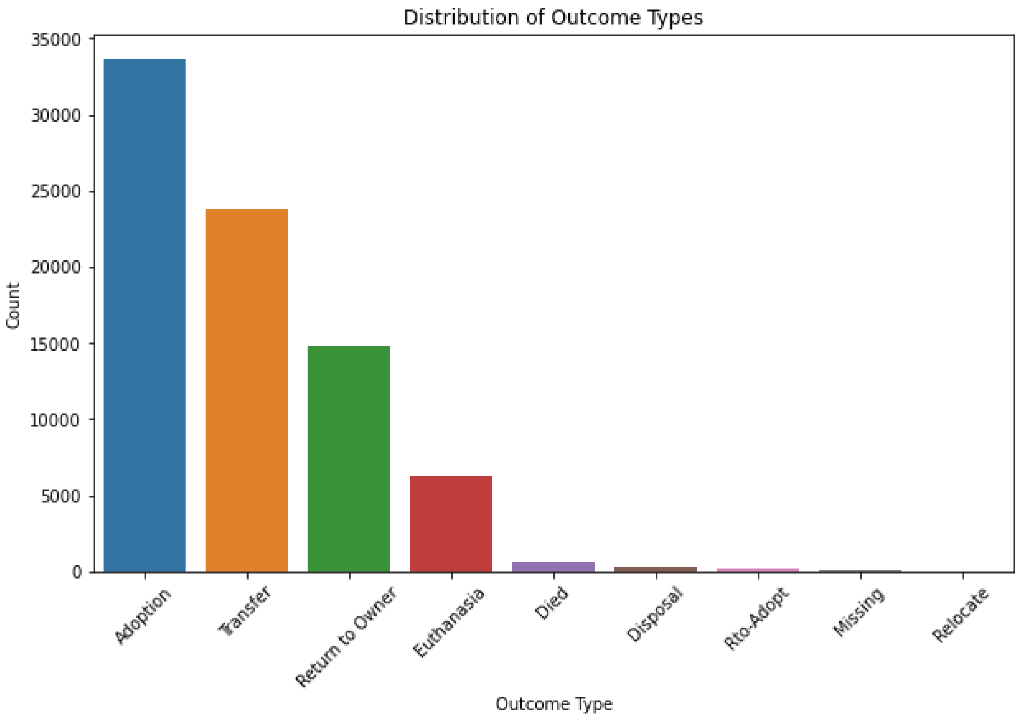
Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a vital step in any data analysis process, as it helps uncover patterns, detect anomalies, and lay the groundwork for further analysis. It's essentially the process of getting to know the data—understanding its structure, spotting inconsistencies, and identifying relationships between variables. EDA is especially important because it allows us to clean and prepare the data properly, ensuring that our models and insights are accurate and meaningful. Visualizations play a key role here, offering a quick and intuitive way to grasp complex information. For example, pie charts can show the proportion of different outcomes like adoptions or transfers, while bar charts highlight the most common animal types entering the shelter. Scatterplots or heatmaps can reveal relationships, like how age might influence an animal's likelihood of adoption. By turning raw numbers into visual stories, EDA and visualizations make it easier to understand the data, form hypotheses, and ultimately guide decision-making with confidence.

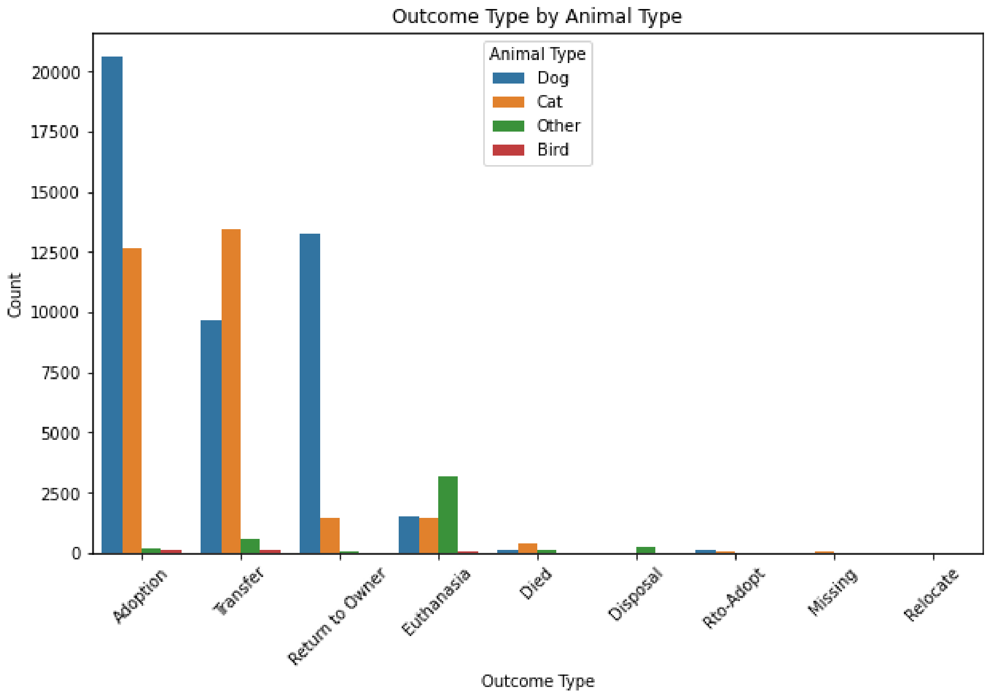
Below are some of the visualizations I used as a part of this project to get some useful insights about the dataset.

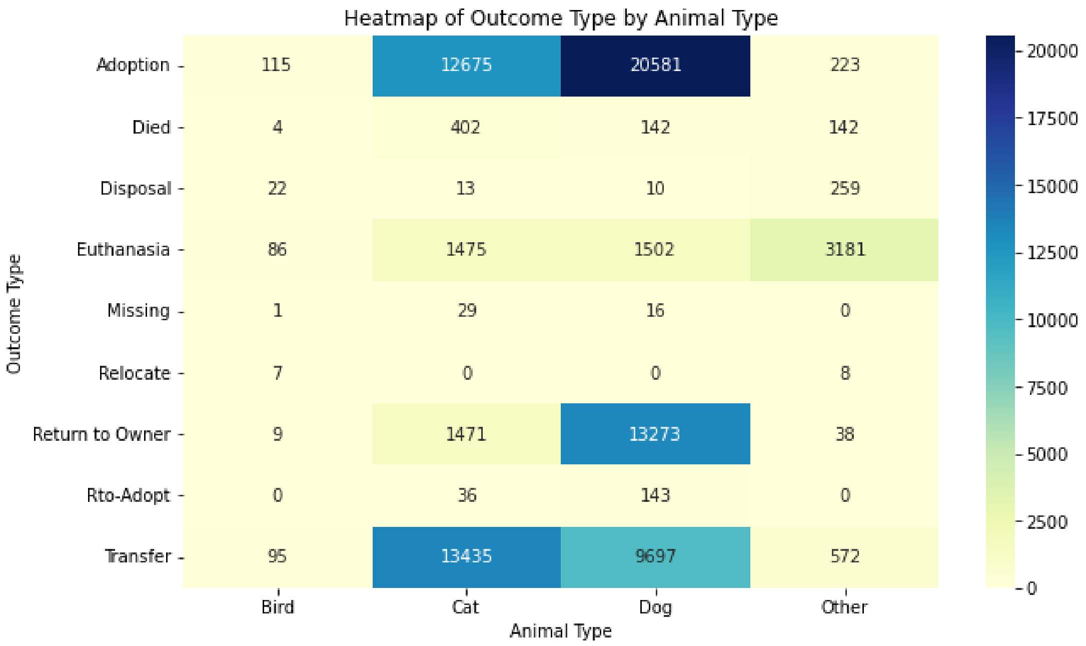


This histogram shows the age distribution of animals. It is evident that younger animals (represented by lower age values) are significantly more likely to be adopted or noticed. The steep decline in frequency as the age increases suggests that the interest in adopting animals diminishes with age, highlighting the importance of prioritizing younger animals in adoption drives.

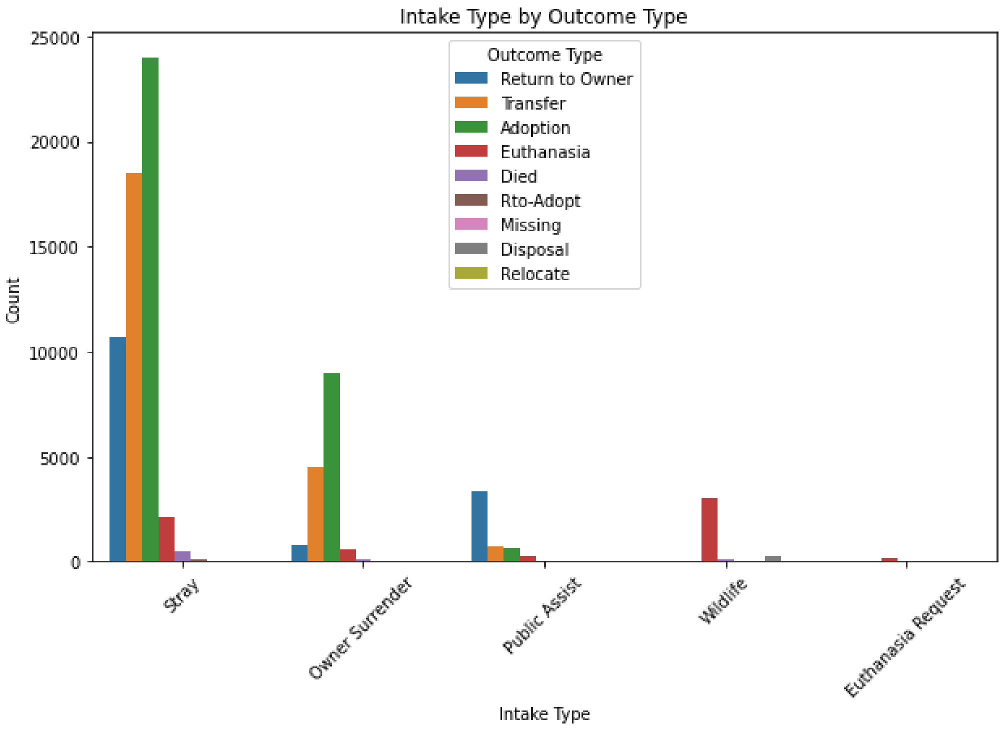


This bar chart reveals that "Adoption" is the most common outcome, followed by "Transfer" and then "Return to Owner." This indicates that shelters are highly successful in finding new homes for animals, with transfers to other facilities also playing a significant role. The relatively lower frequency of reunification with owners suggests a need for improving efforts to connect lost animals with their families. The minimal occurrences of unfavorable outcomes, such as "Euthanasia" and "Died," highlight the shelter's focus on ensuring positive results for animals.





This chart shows how outcomes vary across different animal types. Dogs and cats dominate most categories, with dogs more frequently returned to owners and cats transferred or adopted. Birds and other animals appear less frequently, likely due to lower intake rates or different handling methods. The distribution underscores the need for species-specific strategies in shelters.



This graph visualizes the distribution of various outcome types for animals in the shelter, such as "Return to Owner," "Adoption," "Transfer," and others for different kinds of animal intake categories . "Return to Owner" and "Adoption" stand out as the most frequent outcomes, highlighting successful efforts in reuniting animals with their owners and finding new homes.

However, the notable counts of "Transfer" indicate that many animals are being moved to other facilities or organizations. Less frequent outcomes like "Euthanasia" and "Died" underscore the challenges faced by shelters, particularly for animals with health or behavioral issues. This chart emphasizes the importance of enhancing adoption rates while reducing unfavorable outcomes like euthanasia.

Models and Methods

Before applying any statistical or machine learning techniques, careful data preprocessing was conducted to ensure the dataset was clean, relevant, and efficient for analysis. First, missing values in numerical columns were handled by imputing with mean values, while categorical columns were imputed with their most frequent values. Label encoding was applied to categorical features such as breed, intake_type, and sex_upon_outcome, transforming them into numerical formats that machine learning models can process. To improve model efficiency and avoid computational overhead, feature selection was performed to identify the top 15 most relevant features, reducing the dataset from 41 columns to a concise yet informative subset. This not only saves time but also minimizes the risk of overfitting, ensuring that models focus on the most impactful variables. These preprocessing steps laid the foundation for effective modeling and analysis.

1. Statistical Methods

Statistical techniques provide the foundation for understanding the dataset's structure and relationships:

- **Maximum Likelihood Estimation (MLE)**: MLE is used to estimate the parameters of statistical models, ensuring that the observed data is most probable under the fitted model. This helps in modeling categorical outcomes like adoption or transfer.
- **Ordinary Least Squares (OLS)**: OLS is employed for regression analysis, allowing us to quantify relationships between features like age, intake type, and time in shelter with outcome probabilities.

These methods are valuable for their interpretability, offering insights into how individual features influence outcomes.

2. Machine Learning Models

A range of machine learning models is used to predict the outcomes for animals based on their intake characteristics. Each model contributes unique strengths:

- **Logistic Regression**: This baseline classification model helps understand the probability of specific outcomes, such as adoption versus euthanasia.
- **K-Nearest Neighbors (KNN)**: KNN is a simple yet effective model that predicts outcomes based on the similarity between animals.
- **AdaBoost**: This ensemble method builds a strong classifier by combining multiple weak classifiers, enhancing the predictive accuracy for complex datasets.
- **Gradient Boosting**: Known for its flexibility, Gradient Boosting identifies subtle patterns in the data and improves predictions by focusing on errors made by previous models.

- **XGBoost**: XGBoost, a highly efficient implementation of Gradient Boosting, delivers the best performance in this study. It outperforms other models by optimizing both speed and accuracy, making it the preferred choice for outcome prediction.

The combination of statistical techniques and machine learning models provides a balance between interpretability and predictive power. Statistical methods like OLS help explain relationships between features, while advanced machine learning models like XGBoost and Gradient Boosting offer highly accurate predictions. This layered approach ensures that both insights and predictions are actionable, helping shelter staff make better decisions to improve animal outcomes.

Results

Evaluation of Models for Animal Shelter Outcome Prediction:

Logistic Regression:

- **Accuracy:** 53.7%
- **Confusion Matrix:**
  - This model didn't perform very well overall. It struggled a lot with predicting certain outcomes, like some of the rarer classes (e.g., class 1, class 2, class 7), which is reflected in the low recall and precision for those classes. The model seems to predict the more common classes (like class 0, probably adoption) a lot better, but the rare outcomes were almost completely missed.

Classification Report:

- For most categories, the recall and precision were low. The model did okay with class 0 (likely adoption), but for other categories, it just couldn't get it right. The F1 scores were also on the lower side, showing it couldn't balance precision and recall very well.

K-Nearest Neighbors (KNN):

- **Accuracy:** 55.9%
- **Performance:**
  - The KNN model achieved an accuracy of 55.9%, which is slightly better than the logistic regression and AdaBoost models but still falls short compared to Gradient Boosting and XGBoost. KNN did reasonably well in predicting the more common outcomes, similar to the other models, but struggled with rarer classes.
  - Since KNN relies heavily on distance metrics, its performance might be affected by the presence of noise or irrelevant features. It's also sensitive to the choice of **K** and the **distance metric**, so fine-tuning these parameters could potentially improve performance. However, it wasn't able to outperform the other models in terms of accuracy or class prediction balance.

Suggestions for Improvement: Tuning the K Value: The choice of **K** (number of neighbors) plays a crucial role in KNN's performance. Experimenting with different values could potentially improve accuracy.

AdaBoost:

- **Accuracy:** 52.9%
- **Confusion Matrix:**
  - AdaBoost didn't do much better than logistic regression. It predicted some classes (like class 0) decently, but struggled with the rare classes, much like logistic regression. The model also seemed to favor predicting the more common outcomes, but it wasn't very good at predicting others.

Classification Report:

- Like the logistic regression model, AdaBoost had very low recall for some classes (such as class 1 and class 5), meaning it missed a lot of those outcomes. However, it was able to predict class 3 (maybe transfer) a little better, which is one positive takeaway.

Gradient Boosting:

- **Accuracy:** 82.7%
- **Confusion Matrix:**
  - Gradient Boosting was definitely an improvement over the previous two models. It had much better accuracy and did a solid job with a variety of outcomes, especially class 0 (adoption) and class 3 (transfer). However, it still struggled with some rarer categories, like class 4, showing that there are still areas to improve.

Classification Report:

- Precision and recall were much better than before, especially for class 0 and class 3. But for class 4, the model really didn't perform well, which is likely because that class is underrepresented in the dataset. Still, overall, this model showed a good balance between precision and recall with decent F1 scores.

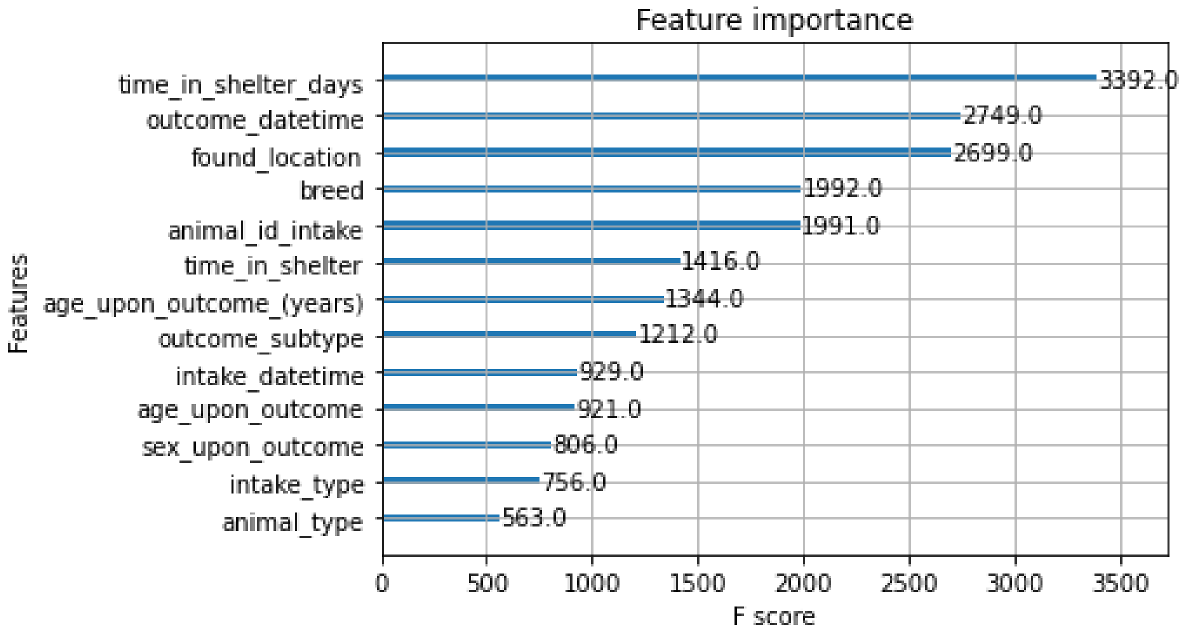
XGBoost:

- **Accuracy:** 84.3%
- **Confusion Matrix:**
  - XGBoost was the best-performing model by far, with the highest accuracy. It did a great job predicting several categories, especially class 0 (adopted) and class 3 (transfer). It even managed to handle some of the rarer categories (like class 8) better than the others.

Classification Report:

- XGBoost had strong performance across the board, with high precision and recall for several categories. For common classes like 0 (adopted) and 3 (transfer), the model nailed it. For rarer classes, it was still hit or miss, but overall, this model was by far the best at predicting the different outcomes.

- Feature importance plot using XGBoost.



Conclusion

In this analysis, we evaluated multiple machine learning models to predict animal shelter outcomes. While **XGBoost** emerged as the top performer with the highest accuracy and the best ability to handle imbalanced classes, **Gradient Boosting** also showed reasonable performance. Despite some challenges with rare class prediction, these models offer useful insights into understanding animal shelter outcomes. Moving forward, improving model performance could involve fine-tuning hyperparameters, balancing the dataset, and exploring additional features or techniques for handling class imbalances.

Future Work

While this analysis provides valuable insights into predicting animal shelter outcomes, there are several avenues for further improvement and exploration:

1. **Hyperparameter Tuning:** Although some initial models performed reasonably well, fine-tuning the hyperparameters for each model could improve their accuracy. Techniques like grid search or randomized search can help identify the optimal settings for better performance.
2. **Feature Engineering:** Further exploration of the dataset might uncover new features or interactions that can improve model performance. Features such as the animal's breed, age, or health status, as well as temporal factors like the time of year, could provide more detailed insights into the likelihood of adoption or euthanasia.
3. **Ensemble Learning:** Combining multiple models in an ensemble approach (e.g., stacking or voting classifiers) could leverage the strengths of different models and improve overall predictive performance.

4. **Model Interpretability:** Further work can focus on making the models more interpretable, especially for stakeholders in the animal shelter industry. Tools like SHAP values or LIME can provide more transparency regarding model predictions and feature importance, which would be valuable in decision-making processes.

By addressing these areas, we could further enhance the accuracy and practical utility of the model, leading to better outcomes for animals in shelters.

Resources

- **Dataset - https://www.kaggle.com/datasets/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes**
- **Reference documentation - https://xgboost.readthedocs.io/en/stable/**