CST338: Data Science 06/14/2024
Final Project Report

SoCalSolutions

Solving The Future

Team:

Ben Denis	bdenis@csumb.edu
Mohe Dean Hajjar	mhajjar@csumb.edu
Rahim Sarpas	rsarpas@csumb.edu
Sandor Faya	sfaya@csumb.edu

GitHub:

https://github.com/mDeanHaj/Pet-Pals-383

Presentation:

SoCalSolutions Presentation

Table of contents

Introduction

This project was undertaken to predict the outcomes of animals in an animal shelter using machine learning techniques. This can help in understanding and improving the decision-making process in animal shelters, potentially leading to better outcomes for the animals.

The research question of the project is to determine if it is possible to accurately predict the outcome type (such as adoption, transfer, return to owner, or euthanasia) for animals in the shelter based on various features such as the animal's type, sex, primary color, secondary color, intake condition, intake type, and age at intake.

The purpose of the project is to try and identify key factors that influence the outcomes of animals in shelters, which could be used to implement strategies for improving these outcomes.

Selection of data

The dataset used in this project is sourced from the City of Long Beach's open data portal, specifically from the "Animal Shelter Intakes and Outcomes" dataset. The URL to access this dataset is:

https://data.longbeach.gov/api/explore/v2.1/catalog/datasets/animal-shelter-intakes-and-outcomes/exports/csv?lang=en&timezone=America%2FLos_Angeles&use_labels=true&delimiter=%2C

The "Animal Shelter Intakes and Outcomes" dataset comprises detailed records of approximately 27,000 animals that have passed through a shelter system. Each entry in the dataset represents an individual animal, providing comprehensive information about its journey through the shelter. The dataset includes information about each animal’s species, DOB, sex, color and health. As well as intake date and outcome.

Even though the dataset was very good overall, several data munging and feature engineering steps were necessary to prepare it for machine learning. Missing values were handled appropriately. Categorical variables like 'Animal Type' and 'Outcome Type' were transformed into numeric values, ensuring compatibility with algorithms. A critical new feature, 'Age at Intake', was created by calculating the difference between 'Intake Date' and 'DOB', providing a

meaningful metric for prediction. Any missing values in this new feature were filled with the mean age to maintain data integrity. Additionally, all features were scaled using StandardScaler to normalize the data, which is essential for optimal model performance. These preprocessing steps ensured that the dataset was clean, consistent, and in a format appropriate for training a logistic regression model.

Methods

The research question regarding animal shelter intakes and outcomes was addressed using a variety of tools that facilitated data manipulation, analysis, modeling, and visualization.

The primary tool used for data handling and manipulation was Python, using Pandas for efficient data loading, cleaning, and feature engineering. NumPy also played an important role handling numerical operations underlying Pandas operations.

For preprocessing categorical data, **Scikit-learn (sklearn)** played a pivotal role with **OrdinalEncoder** to transform categorical variables ('Animal Type', 'Sex', 'Primary Color', etc.) into numerical values suitable for machine learning models. **StandardScaler** from Scikit-learn standardized numeric features like 'Age at Intake' to ensure uniformity and optimize model performance.

The analysis was powered by **Seaborn** and **Matplotlib**, which enabled insightful visualizations. **Seaborn** was employed for creating informative visual summaries such as count plots ('Distribution of Animal Types', 'Distribution of Intake Conditions') and box plots ('Age at Intake vs Outcome Type'). These visualizations provided crucial insights into the distribution of animals across types and intake conditions, as well as the relationship between age at intake and the eventual outcome.

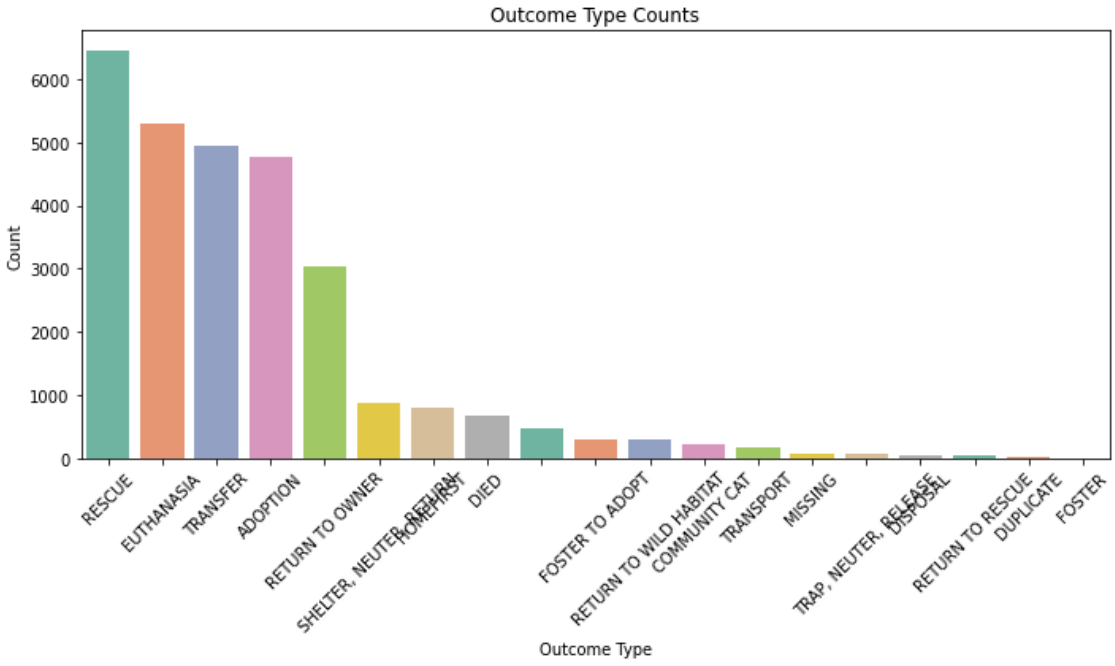
The core analytical tool used to model and predict outcomes was **LogisticRegression** from Scikit-learn, configured to handle the prediction of 'Outcome Type' based on the transformed and scaled features.

Overall, the combination of Python, Pandas, NumPy, Scikit-learn, Seaborn, and Matplotlib worked seamlessly together to provide a clearer understanding of how certain variables influence shelter outcomes.

Results

The primary goal of this project was to predict the likelihood of pet adoption using a logistic regression model. We evaluated the model's performance using accuracy as the metric. After training the model on the scaled data, the logistic regression model achieved an accuracy of 28.59%. Based on the analysis conducted on the dataset from the animal shelter, several key findings have emerged that address the research question regarding outcomes for animals. The study focused primarily on the most prevalent outcomes—adoption, rescue, transfer, return to owner and euthanasia—while filtering out less frequent outcomes to ensure accuracy in addressing our research objectives. These insights can help shelters understand the predominant types of animals and characteristics they need to focus on for adoption campaigns.

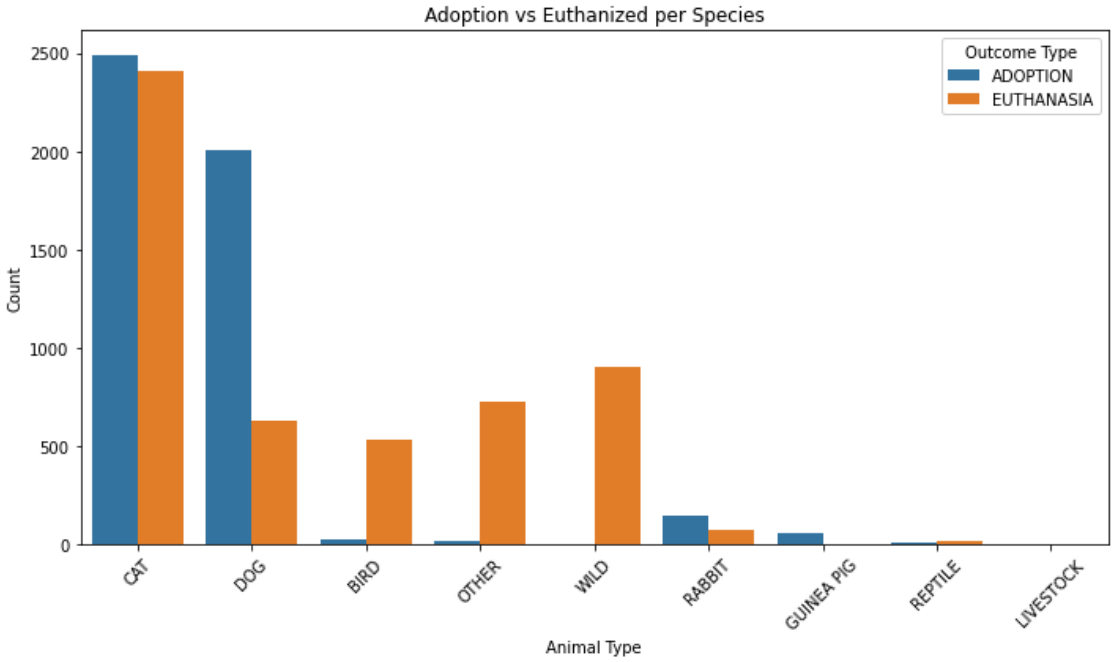
Among the outcomes studied, rescue emerged as the most frequent outcome, closely followed by euthanasia.

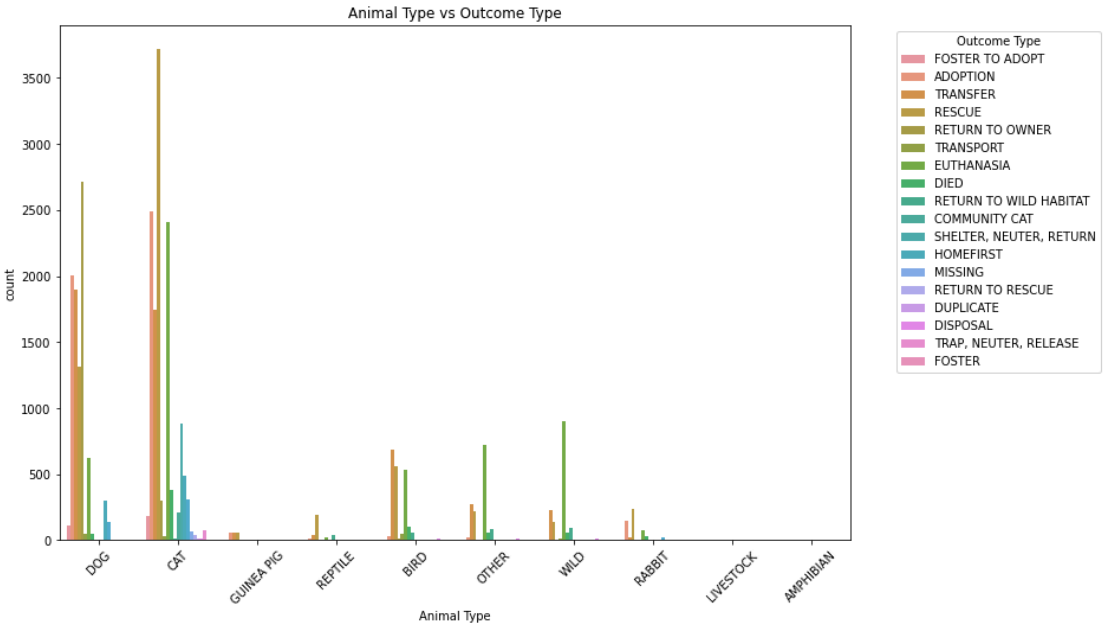


This indicates a significant pattern within the shelter's operations, where animals are frequently saved or, unfortunately, euthanized. Notably, the data shows a positive trend where animals are more likely to be rescued or adopted rather than euthanized, reflecting a reassuring aspect of the shelter's impact.

Analyzing outcomes by species reveals an interesting trend: cats tend to have a higher likelihood of being euthanized compared to dogs. However, both cats and dogs show a predominant likelihood of being adopted or rescued, underscoring their appeal and the shelter's success in rehoming these animals.

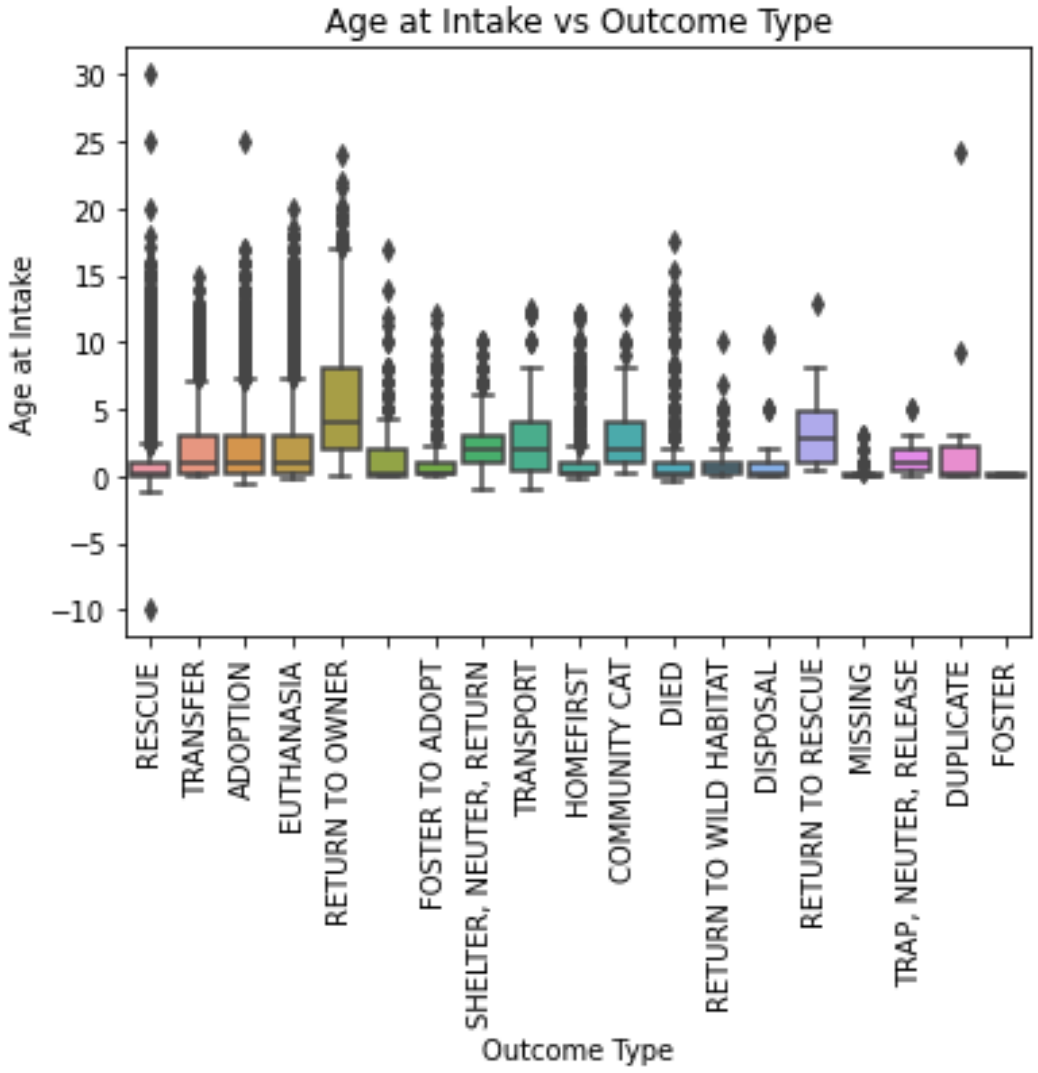
Unfortunately, this is not the case for birds and wildlife animals. Birds and wildlife animals are noted to have a high incidence of euthanasia, reflecting the challenges shelters face in rehabilitating and rehoming these species.



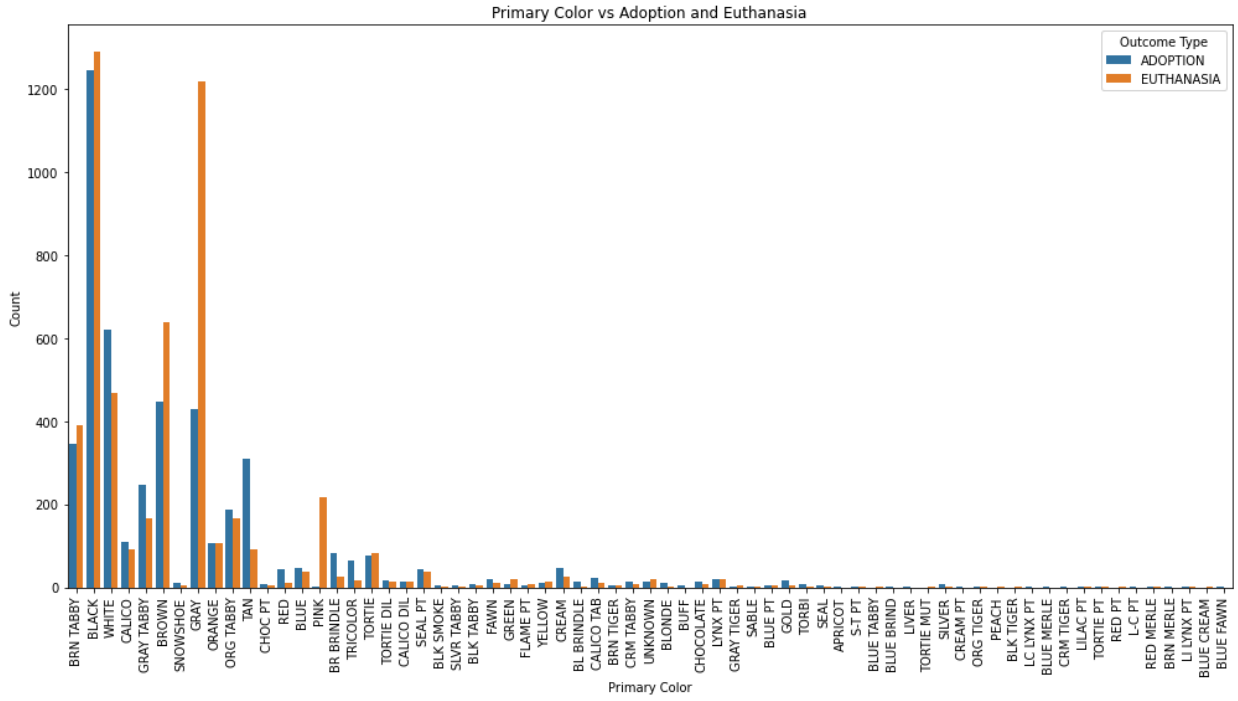


Further analysis by age group highlights that younger animals, particularly those between 0-3 years old, are significantly more likely to be adopted. This age group experiences the highest rates of adoption, emphasizing the preference for younger pets among potential adopters.

The dataset also reveals that pets aged 5 years and older are more frequently returned to their owners rather than adopted. This finding suggests a shift in adoption preferences towards younger animals and a higher incidence of returning older pets to their original homes.

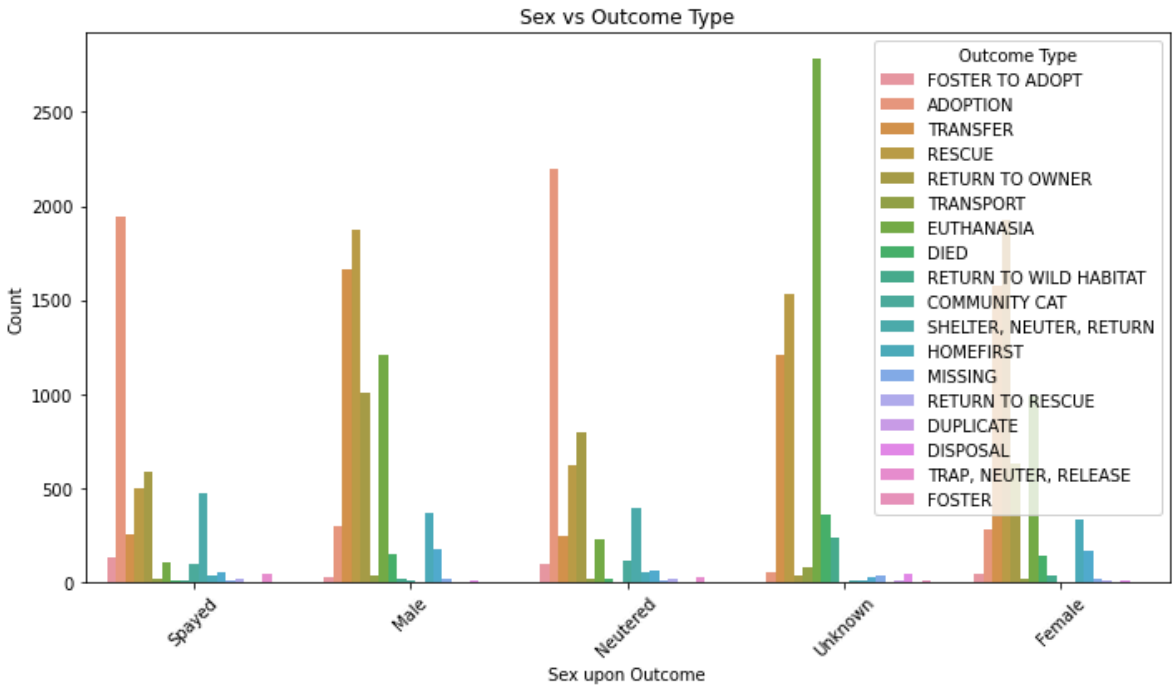


Color also plays a surprising role in adoption rates, with certain colors such as gray and pink being associated with a higher likelihood of euthanasia. In contrast, animals of white, tan, and cream colors exhibit higher rates of adoption and rescue, indicating color preference among adopters.

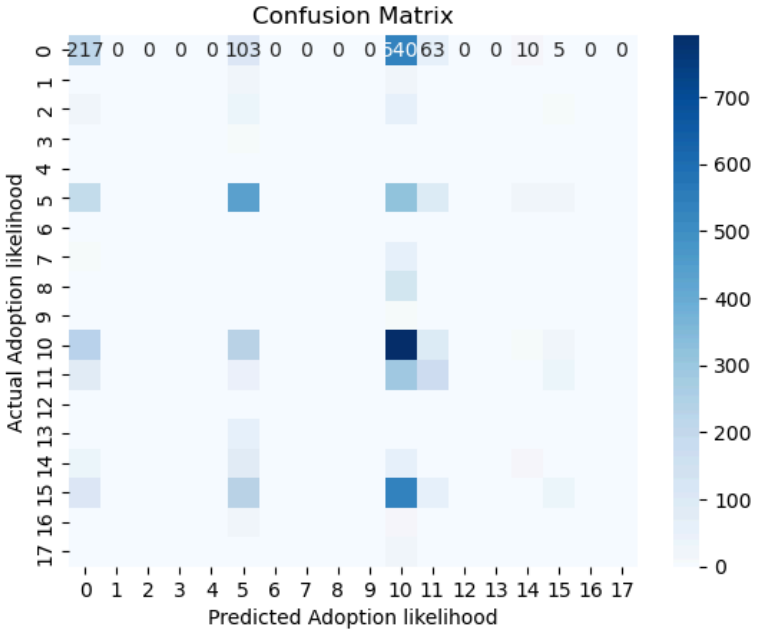


Notably, the data highlights the significant impact of spaying or neutering on adoption rates, with spayed/neutered animals showing adoption rates as high as 75%. This emphasizes the importance of these procedures in increasing the chances of adoption for shelter animals.

Sex also influences outcomes, with male animals having a higher likelihood of euthanasia compared to females, who are more frequently adopted or rescued. Additionally, animals whose sex is unknown tend to face a higher incidence of euthanasia, suggesting a need for improved identification and care protocols for these animals.



With our preliminary work on machine learning, we made a model to split the dataset into a training and testing set for itself. The goal of this was to increase the accuracy of predicting the likelihood of adoption to the actual likelihood of adoption.



Needless to say, the model’s accuracy was a given 29% which shows that there is more room for improvement than was suggested. The confusion matrix revealed the model's performance in predicting each outcome type. While the model had some success in correctly predicting adoptions, it struggled with other outcome types, highlighting the need for further refinement and potential inclusion of additional features to improve prediction accuracy.

Discussion

The findings of this study highlight several important implications and areas for future research. The ability to predict adoption likelihood based on various attributes of the animals can significantly enhance the efficiency and effectiveness of shelter operations. By identifying animals less likely to be adopted, shelters can implement targeted interventions such as increased visibility, special adoption events, or additional support. This strategic approach can help shelters optimize their resources by focusing efforts on animals with lower adoption probabilities, thereby improving overall adoption rates and reducing shelter overcrowding.

Our findings align with existing research indicating that younger animals and certain breeds or types are more likely to be adopted. However, this study provides a more comprehensive analysis by considering a wider range of attributes and utilizing machine learning techniques for prediction. Previous studies have often focused on individual factors influencing adoption, whereas our model integrates multiple variables, offering a holistic view of the factors affecting adoption outcomes. This integrated approach allows for a deeper understanding of the complex interplay between various attributes and their impact on adoption success. Such is the case with the identification that spaying or neutering the animals, or implementing improved identification and care protocols, significantly increases the chances for adoption.

Future research should explore the inclusion of additional features, such as behavioral assessments, medical history, and more detailed intake conditions, to enhance the predictive power of the model. Investigating other machine learning algorithms, such as decision trees, random forests, or neural networks, could potentially improve prediction accuracy.

This ongoing research will be crucial in refining predictive models and improving outcomes for animals in shelters, ultimately leading to more effective and humane animal welfare practices.

Summary

In conclusion, the findings on this project provide valuable insights into the factors influencing outcomes for shelter animals, highlighting successful adoption and rescue rates for certain species, colors, and conditions, while also pointing to areas where interventions could potentially improve outcomes for others. The visualizations created from the data effectively illustrate these trends, providing a clear picture of the shelter's operations and the factors impacting the lives of animals in its care. These insights can guide shelters in optimizing their practices to increase adoption rates and improve the overall well-being of the animals.

The key findings of this project include the logistic regression model's achievement of a 28.59% accuracy in predicting adoption outcomes. This relatively modest accuracy underscores the complexity of predicting adoption outcomes but also highlights areas for model improvement.

Overall, this project demonstrates the potential of using machine learning techniques to predict pet adoption likelihood and offers valuable insights for improving shelter outcomes. Future research and refinement of the model can further enhance its predictive power and practical application in animal shelters. By integrating more features and exploring different algorithms, the predictive accuracy can be improved, leading to better resource allocation and more effective interventions.

References

- City of Long Beach Open Data Portal. (n.d.). Animal Shelter Intakes and Outcomes Dataset. Retrieved from [Long Beach Data Portal](https://data.longbeach.gov/api/explore/v2.1/catalog/datasets/animal-shelter-intakes-and-outcomes/exports/csv?lang=en&timezone=America%2FLos_Angeles&use_labels=true&delimiter=%2C)

- Scikit-Learn: Machine Learning in Python. (n.d.). Retrieved from [Scikit-Learn](https://scikit-learn.org/stable/)

- Seaborn: Statistical Data Visualization. (n.d.). Retrieved from [Seaborn](https://seaborn.pydata.org/)

- Matplotlib: Visualization with Python. (n.d.). Retrieved from [Matplotlib](https://matplotlib.org/)