Outcome Prediction for Shelter Dogs and Cats

Abhigna Sowgandhika Vadlamudi
Department of Computer Science
University of Houston
2268166

Ahmed Hussain Syed
Department of Computer Science
University of Houston
2273473

Abstract

With the ongoing demand for data analysis and prediction, machine learning has become a pivotal area of research and experiment. This project outlines a data-driven approach to predict the outcomes of shelter dogs and cats, aiming to enhance resource allocation, improve animal welfare, and reduce euthanasia rates. While most of the state-of-the-art approaches prioritized leveraging a single machine learning algorithm, our research emphasizes integrating various features including demographic information, behavioral traits, and health status which enables the proposed framework to improve upon existing predictive accuracy. The project will initially focus on a localized open dataset from the City of Austin. Predictive analysis is performed using Naive Bayes, K-nearest Neighbours, Classification Tree, and Random Forest. Evaluation will be conducted using historical shelter data, with accuracy, precision, and recall metrics. By providing shelters and rescues with actionable insights, the project aims to facilitate quicker outcomes for animals, strengthen partnerships with fosters, and ultimately improve the well-being of shelter animals across the country.

1 Introduction

Within the past few decades, Machine Learning (ML) has been used in an attempt to analyze and interpret patterns found in data across numerous fields of technology. Recommendation engines, facial recognition, and spam filtering are some of the most popular applications of machine learning in real-life situations to aid humans in their daily tasks. Since the 1940s, Machine Learning concepts have been leveraged for analyzing information and making future predictions. This is what we aim to implement in the current project as well. The Shelter Animals Count organization, estimates in their national database that 6.5 million cats and dogs entered shelters and rescues across America in 2023. Of these, 4.8 million were adopted, 690,000 were euthanized and the rest were still waiting for an outcome. This data is important because it helps us to understand the accommodation and resource constraints shelters have to face while handling increasing numbers of intakes. In Machine Learning, predictive analytics models are created to evaluate past data, uncover patterns, analyze trends, and leverage that insight for forecasting future trends. These models contain basic steps that are implemented always: identifying the problem, building the model, testing, evaluating uncertainty. Weather forecasting, risk modeling, diagnosis of healthcare, fraud detection, and supply chain management all now make use of predictive models to cater to the needs of investors and consumers. Our work implements predictive analysis in a way that the results will provide detailed insights into the leading factors of whether an animal will be adopted or not to inform the operations of the Austin Animal Shelter.

Number of	Count
Animals Counted	26.6M
Participating Organisations	7026
Data Points Collected	172
Reporting States	59

Table 1: Animal Sheltering Statistics by Shelter Animals Count Organisation

2 Prior Work

Up until this point, there have been a few notable papers that have introduced new architectures to the field, based on other ML concepts. Previous research in animal shelter management has primarily focused on descriptive analysis and simple statistical models to understand adoption patterns and euthanasia rates. Baseline approaches often involve basic demographic features of animals such as age, breed, and gender, along with shelter-specific variables like intake location and time spent in the shelter. While these approaches provide some insights, they often lack predictive accuracy and fail to capture the complexity of factors influencing animal outcomes.

(Sazara and Gao, 2022)
(Janae and Suchithra, 2021)

3 Problem Statement

In this project, the aim is to assess how comparing the accuracies of different Machine Learning models, specifically K-nearest neighbors, Naive Bayes, and Random Forest, will help us build a proper framework to achieve a prediction model. We also perform exploratory analysis on the datasets and factor out the dominant features.

4 Methodology

Originally, the intention was to run the datasets through pre-existing predictive models, identify their respective underperforming components, and build a new model addressing these issues. This was changed for two reasons: the initial plan necessitated a large model that was infeasible to train on a local machine, and access to pre-existing frameworks was limited due to the technology used by them being redundant. For these reasons, we decided to go with a comparative approach on simpler datasets, which could be accomplished in the current time frame.

4.1 Data

The datasets used in our current work have been obtained from the City of Austin's `open data portal.`[1]

- The intake dataset contains 160k rows and 12 columns which include the animalid, name, found location, intake type, condition, breed, etc.

- The outcome dataset also contains the same number of rows and columns where each row represents one outcome per animal per encounter.

We first preprocess this data and perform a thorough analysis. Seventy percent of the available data has been used to train the models while the remaining thirty percent has been used for testing.

4.2 Technical Approach

Building upon existing research, we explored novel directions in predictive modeling for shelter animal outcomes by integrating a wide range of features including not only demographic information but also behavioral traits, health status, and historical shelter data. Additionally, we leveraged advanced machine learning algorithms such as classification trees, random forests, etc., to capture complex relationships among predictors and predict adoption likelihood more accurately. Furthermore, we also investigated the potential of natural language processing techniques to analyze textual data such as animal descriptions and adoption narratives, to extract valuable insights to inform predictive models.

4.3 Exploratory Analysis

The first use of the available data was during preprocessing. We cleaned each dataset, removed redundant entries, and analyzed the data features and their significance.

> "Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods."

This was followed by feature extraction which played a prime role in helping us understand what metrics to consider while passing the data through Machine Learning models.

- Majority of newly intake cats and dogs were between 1 and 3 years and most of the remaining were young, under 6 months. There was also a small percentage of adult animals with ages more than 7 years.

- Comparing the intake and outcome entries of animals also determined that most of the animals spent under 7 days in the shelters with approximately 300 staying for more than 2 years.

- 77 animals from the outcome dataset were euthanized as per records.

- To encourage the benefits of neutering and spaying, 27139 animals were operated on thereby reducing the rate of reproduction.

[1] https://data.austintexas.gov/

4.4 Visual Analysis

Representing the data in the form of graphs and plots is necessary to efficiently summarise large datasets.

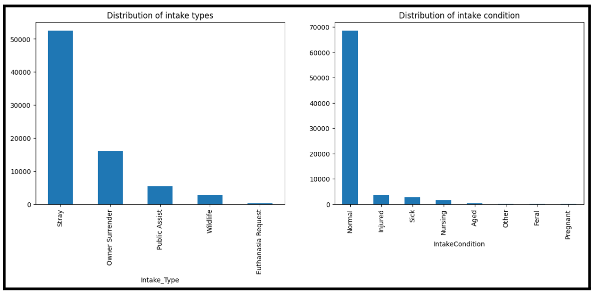

Figure 1: A histogram displaying the types of intake and the conditions of animals upon intake

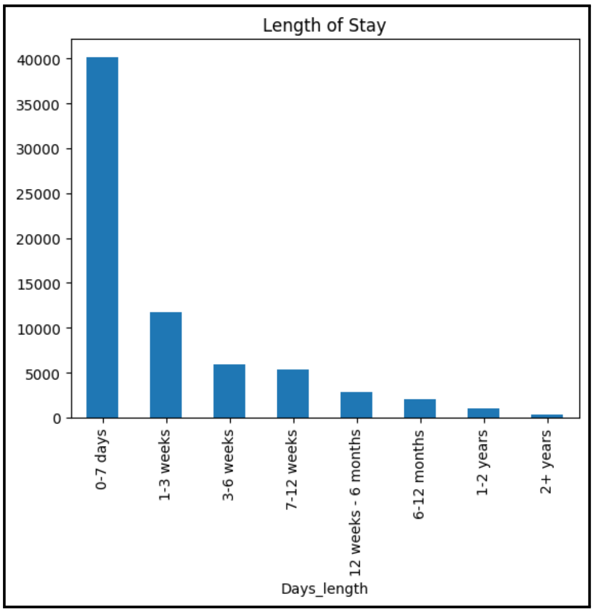

Figure 2: Average length of stay of all animals in the shelter

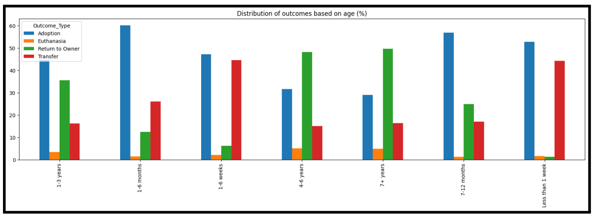

Figure 3: Percentage of animals for each outcome based on categories of age

4.5 Machine Learning Models

After preprocessing and analyzing the data, we start evaluating it by passing it through predictive machine-learning models. Appropriate performance metrics (here, accuracy) tailored to the task of predicting shelter animal outcomes will provide insights into the models' ability to correctly classify animals into their respective categories.

The current task involves binary classification (adopted/not adopted). Thus, we start by setting the target column for prediction and determining the baseline prediction accuracy.

```
majority = float(records['Target'].
value_counts()[0])
total = records['Target'].value_
counts().sum()
baseline = majority/total
print (format(baseline, '.4f') + '%')
records['Target'].value_counts()
```

This determines the baseline prediction accuracy to be 0.5790 percent.

We then split the available data into training and testing datasets:

```
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=
0.3, random_state=1)
```

4.5.1 K-Nearest Neighbours

This algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

We start by assigning a class label based on a majority vote—i.e. the label that is most frequently represented around a given data point is used.

The model then predicts the outcomes of animals by giving the training and testing accuracies as outputs.

```
Train accuracy: 0.6501461227431042
Test accuracy: 0.654570080799589
```

Then, we test for the accuracies by changing the number of neighbors the model is allowed to take as input.

- Number of neighbors = 50

```
Train accuracy: 0.6501060891148565
Test accuracy: 0.654196441081687
```

- Number of neighbors = 250

```
Train accuracy: 0.653869250170143
Test accuracy: 0.6572322637896408
```

- Number of neighbors = 500

```
Train accuracy: 0.6543696705232395
Test accuracy: 0.6582597730138714
```

4.5.2 Naive Bayes Model

This is a supervised machine learning algorithm that is used for classification tasks such as text classification.

According to IBM, "Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes".

We use multinomial Naive Bayes algorithm for our data so that it allows us to use the discrete data present in our datasets. The accuracies obtained are as follows:

```
Train accuracy: 0.646202810360703
Test accuracy: 0.6520480127037505
```

4.5.3 Random Forest Model

This model combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility enable it to handle both classification and regression problems.

Since the random forest model is made up of multiple decision trees, let us first take a look at what can make up individual decision trees. Decision trees start with a basic question, such as, “Do more dogs enter shelters than cats?” Basically, we can form a series of questions to determine an answer, such as, “Which skin color of animals is more preferred among humans?” or “Are puppies more likely to get adopted?”. These questions make up the decision nodes in the tree, acting as a means to split the data.

For our current approach, we are considering a more generalized question, "Will this animal get adopted?" Through this model, we obtain the following accuracies:

```
Train accuracy: 0.6519676528283759
Test accuracy: 0.6520480127037505
```

5 Results

We will now compare the performance of the proposed models against each other to help identify the most effective algorithm for predicting the adoption of animals from a shelter. Based on this, we can quantify improvements in prediction accuracy and computational efficiency. A tabular representation of these model performances can be observed as given below: Based on these findings, we therefore determine the most effective predictive model for our job of predicting whether an animal would be adopted or not. With a test accuracy of 65.82 percent, this model outperforms change by more than 10 percent in the predicting job. 10 percent of 70,000 cases is a significant amount; this will help the Austin Animal Shelter forecast adoption rates for an additional 7,000 animals more precisely.

Model	Test Accuracy
K-nearest neighbors	65.82
Naive Bayes	65.20
Random Forest	65.20

Table 2: Comparison of accuracies given by each ML model

From the exploratory analysis, we also got the following summarised results. The datasets we obtained range from October of 2013 to April of 2017 (see Figure 5 for a distribution of month and year). Overall, most adoptions occurred during Saturdays and Sundays, with 21.6 percent and 20.1 percent respectively (see Figure 6). The most popular month to adopt was July, with 10.5 percent of all adoptions (see Figure 7). There seems to be some seasonality when it comes to adoptions. Holiday months, such as December and January are the next most popular months for adoptions, with 10 percent and 9.1 percent respectively.

6 Discussions

Predicting whether or not an animal will be adopted is a powerful tool. The Austin Animal Shelter might more effectively prioritize animals with little chance of adoption by using our methodology when it comes to putting these animals on their website, asking for transfers, and establishing foster programs. Austin Animal Shelter should attempt giving the unidentified animals names in order to see if this improves adoption rates, as it is well known that an animal's name is a major predictor of adoption. Finally, when room is limited, Austin Animal Shelter can use the results of our exploratory study to prioritize intakes by gender, age, and breed.

7 Limitations

We also tested for a classification tree model which gave a better test accuracy (69.8 percent), but it was unpruned, making it nearly hard to utilize in a real-world setting. The values included in the leaf nodes, which form the base of the classification tree, showed the relative results of our predictions. The Intake Type was the feature that was used the most, as the classification tree illustrates. The likelihood of adoption is doubled for animals that were owner surrendered, but not for those that weren't 0.66 times more likely to be embraced.

Though we present a novel approach to predictive modeling for shelter animal outcomes, several limitations exist that warrant consideration. Firstly, the study lacks a comprehensive comparison of algorithms, potentially limiting insights into the relative performance of the chosen methods. Secondly, due to time and resource constraints, we did not delve into the process of feature selection and engineering, which could impact model performance and interpretability. Thirdly, the issue of data imbalance, common in shelter animal datasets, was not addressed, raising concerns about biased model evaluations and predictions. Furthermore, there is limited discussion on model interpretability, hindering stakeholders' ability to understand and trust the predictions.

Since we relied on data from just one single source, generalizability is also questionable. Addressing these limitations would enhance the robustness, reproducibility, and applicability of the research findings in real-world shelter settings.

References

Bradley Janae and Rajendran Suchithra. 2021. Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation. *BMC Veterinary Research*, 17(70).

C. Sazara and X. Gao. 2022. Predicting animal shelter pet adoption times and feature im- portance analysis using catboost. In *2022 IEEE 11th International Conference on Intelligent Systems (IS)*.