COMPSCI 526 Milestone Report

Title: Data-Driven Discoveries in Pet Adoption Patterns

Project Members:
Yushan Shi ys468
Yuqi Su ys457
Yurui Wu yw695
Mingxin Liu ml754

Youtube Link: https://youtu.be/a4Gq80cDGcc
Github Repo Link: https://github.com/ArikaGrimmer/CS526TeamProject

Introduction

Our project focuses on the critical issue of pet adoption speed, aiming to uncover the factors that influence how quickly animals are adopted. With millions of stray animals facing euthanasia or homelessness each day, understanding these factors is essential for improving adoption rates and promoting global animal welfare. We are using the PetFinder.my Adoption Prediction dataset from Kaggle, which provides detailed profiles of pets available for adoption. By analyzing characteristics such as breed, age, health status, and multimedia elements like photos and videos, we seek to identify patterns that lead to faster adoptions.

Through exploratory data analysis (EDA), we will examine relationships between key features and adoption speed. Using advanced techniques like SHAP (SHapley Additive exPlanations) values, we aim to understand how these characteristics interact to influence adoption outcomes. Our goal is to offer shelters actionable insights, helping them create more effective pet profiles and ultimately increasing adoption rates.

Dataset Description

We utilized the PetFinder.my Adoption Prediction dataset from Kaggle, which provides detailed information about each pet profile, including basic characteristics, photos, and textual descriptions of the pets. The characteristic features and textual descriptions are contained in a single training file, while the pet photos are stored separately and labeled by profile ID.
Since analyzing photos and descriptions in depth would require advanced techniques like deep learning, we mainly focused on the characteristic features, which can be divided into categorical, ordinal, or numerical ones.

Categorical features include pet name, type, gender, breed, color, and deworming, sterilization, and vaccination status. Ordinal features include maturity size (small, medium, etc.), fur length, and health condition. Numerical features include pet age, quantity (number of pets in a profile), adoption fee, video count, and photo count.

To incorporate information from the textual descriptions, a preliminary sentiment analysis was conducted on each sentence in pet description using Google's Natural Language API. This allowed us to quantify the overall sentiment of each description, offering a simpler alternative by translating the textual content into numerical sentiment scores. Both positive and negative sentiment scores were provided separately, allowing us to distinguish descriptions with mixed

sentiment from those with neutral tones.

Two versions of the dataset—with and without sentiment scores—were processed for further exploration. Textual columns, such as pet names and IDs, were removed, and the data was split into training and validation sets. We normalized all numerical columns and filled any missing values using the mean.

Analysis

Exploratory Data Analysis (EDA): we first focus on examining the relationships between key categorical features and `AdoptionSpeed`. This analysis included visualizations such as bar charts, pie charts, line graphs, box plots, and heatmaps. Given the differing adoption patterns between cats and dogs, we stratified these visualizations by animal type. Additionally, due to the categorical nature of the features, we employed the Chi-square test to assess the statistical significance of these relationships.

Ensemble model analysis: Apart from traditional statistical analysis, we also took another approach to examine the potential relationship between the adoption speed and the features of the pets. To do so, we implemented a multiclass XGBoost model to predict the adoption speed. To predict multi labels, we set objective function and eval_metric as **multi: softmax** and **mlogloss** respectively. After evaluation, we examined the feature importance of the model to explore what could be the potential factors affecting our target variable. We also tuned the hyperparameters by using random search to select 100 combinations of hyperparameters, including n_estimators, max_depth, learning rate, etc. To make our project more flexible, we included the option of using sentimental features. To gain deeper insights into feature interactions and their impact on model predictions, we implemented SHAP (SHapley Additive exPlanations) analysis. The SHAP analysis, as shown in the figure in the next section, reveals complex interactions between key features such as Type, Gender, Age, and Breed characteristics. The violin plots demonstrate how these features interact with each other and their individual contributions to the model's predictions. Particularly interesting are the symmetrical distributions seen in Gender interactions, suggesting balanced influence across categories, while Age shows more varied interactions with other features, indicating its complex relationship with adoption speed. The Breed1 and Breed2 interactions reveal distinct patterns that could indicate certain breed combinations having significant impact on adoption outcomes. This analysis provides valuable insights beyond simple feature importance rankings, helping us understand how different pet characteristics work together to influence adoption speed.

Narrative & Insights

Exploratory Data Analysis (EDA): By studying relationships between key features and adoption rate, we can see that for both dogs and cats, adoption speed generally decreases as the age of the pet increases, with dogs experiencing a more significant decrease in adoption speed compared to cats as they get older. Also, regarding the health of dogs and cats, being vaccinated and being dewormed correlates with higher adoption speed, as well as a healthier status. All these basic relationships agree with our intuition.

Ensemble model analysis: From our XGBoost model analysis, we gained valuable insights into pet adoption speed patterns through both feature importance and interaction analysis. The SHAP interaction plot reveals complex relationships between pet characteristics,

with notable patterns between Type and Breed1, suggesting certain pet type-breed combinations significantly influence adoption speed. Gender shows balanced interaction patterns across all features, while Age demonstrates varied effects depending on its combination with other characteristics, particularly with Type and Breed1.

The feature importance analysis identifies Quantity and Fee as the most influential factors, both scoring around 0.058, followed by VideoAmt, Age, and PhotoAmt. This hierarchy suggests that listing characteristics (such as batch listings and pricing) have a stronger impact on adoption speed than the pet's inherent characteristics. The significant impact of multimedia content (videos and photos) in the top five features indicates that better visual representation of pets in listings plays a crucial role in the adoption process, potentially serving as a key strategy for improving adoption outcomes.

Timeline

For the next two tasks given by the project, the tentative schedule can be found below:

a. *Team Project Presentation (Oct 25th - Nov 18th):*
    i. *Wrapping up the analysis (Oct 25th - Nov 6th)*
    ii. *Preparing presentation slides (Nov 7th - Nov 13th)*
    iii. *Practicing presentation (Nov 14th - Nov 18th)*
b. *Final Report + Reproducible Code (Nov 22nd - Dec 10th)*
    i. *Polishing codebases (Nov 22nd - Nov 26th)*
    ii. *Drafting the report (Nov 27th - Dec 10th)*

The plans are subject to change but the deadlines of the key tasks will be met.

Contributions

In this project, each team member played a vital role in ensuring a comprehensive analysis. Yuqi led the introduction, setting the context and objectives. Yuri and Yushan collaboratively described the dataset, providing insights into data attributes and structure. Mingxin and Yushan focused on data analysis, using advanced techniques to uncover patterns and trends. Narrative and insights were co-developed by Mingxin, Yuqi, and Yuri to highlight key findings and their implications. Mingxin also managed the project timeline, ensuring timely progress. The final project video was co-created by Yushan and Yuqi, offering a clear presentation of our results and conclusions.