Data-Driven Discoveries in Pet Adoption Patterns

Abstract

This study investigates pet adoption patterns using the PetFinder.my dataset, employing interpretable machine learning techniques to identify factors influencing adoption speed. Key findings highlight the critical role of multimedia and pet attributes like age and breed, offering actionable insights for shelters to enhance adoption rates.

1 Motivation and Problem Statement

The adoption of shelter animals represents a pressing societal challenge, as millions of pets enter animal shelters each year. The American Society for the Prevention of Cruelty to Animals (ASPCA) estimates that approximately 6.3 million companion animals are received by shelters across the United States annually, yet only 4.1 million find homes through adoption. Tragically, this gap results in the euthanasia of roughly 920,000 shelter animals each year. Enhancing the speed of adoptions not only improves the welfare of these animals but also mitigates significant operational pressures on shelters, including overcrowding, constrained budgets, and limited staffing resources. Given the profound societal and economic implications, identifying the factors that influence adoption speed is of utmost importance.

This project is driven by the imperative to uncover actionable insights that can boost adoption rates and equip shelters with data-driven strategies to optimize their operations and outcomes. By employing traditional and interpretable machine learning techniques, the study ensures computational efficiency while delivering meaningful and practical findings. Central to this effort is the critical question of how specific pet attributes affect adoption speed—an understanding that holds the key to improving shelter adoption rates and addressing this multifaceted challenge.

2 Dataset Description

We leveraged the PetFinder.my Adoption Prediction dataset, sourced from Kaggle, which provides comprehensive details on pet profiles, encompassing basic characteristics, photographs, and textual descriptions. The dataset consolidates characteristic features and textual descriptions within a single training file, while pet photographs are stored separately, indexed by profile ID. Due to the computational complexity associated with analyzing images and textual data, our analysis primarily focused on characteristic features, classified into three types: categorical (e.g., type, breed, color, vaccination status), ordinal (e.g., maturity size, fur length, health condition), and numerical (e.g., age, quantity, adoption fee, video/photo count).

To incorporate the textual descriptions, we performed a preliminary sentiment analysis using Google's Natural Language API. This process converted the textual content into numerical scores representing positive and negative sentiment, effectively capturing mixed and neutral tones. Consequently, we prepared two versions of the dataset—one including sentiment scores and one without—for subsequent analysis. Textual columns, such as pet names and IDs, were excluded from both versions. Numerical features were normalized to ensure consistency, and missing values were imputed using the mean of each respective column.

Furthermore, we excluded profiles representing multiple pets due to inconsistencies in feature referencing and frequent discrepancies between features and descriptions, which risked compromising data integrity. This refinement reduced the dataset to approximately 10,000 profiles, yielding a robust sample size sufficient for reliable modeling and statistical analysis.

3 Data Exploration and Analysis

For the initial data exploration, I conducted Exploratory Data Analysis (EDA) to investigate the relationships between critical features—such as age and health status—and adoption speed. Although we initially hypothesized distinct adoption patterns for cats and dogs, the EDA revealed strikingly similar trends across both species. Based on this finding, we consolidated the data for both species in subsequent machine learning analyses, streamlining the process while preserving the integrity and reliability of the results.

3.1 Exploratory Data Analysis

By analyzing the relationships between key characteristics and adoption rates, we observe a common trend for both dogs and cats: adoption speed tends to decline as the pet's age increases. However, the overall pattern remains consistent between the two species. Additionally, when considering pet health, we find that vaccination and deworming are associated with faster adoption rates and generally better health conditions. These fundamental relationships align with our expectations.

To further investigate the potential impact of age imbalance, we analyzed the age distribution and its relationship with adoption speed. The results, presented in Figure 1, indicate that younger animals are more prevalent across all adoption speed categories. Moreover, we do not observe significant differences in age distribution patterns between cats and dogs in relation to adoption speed.

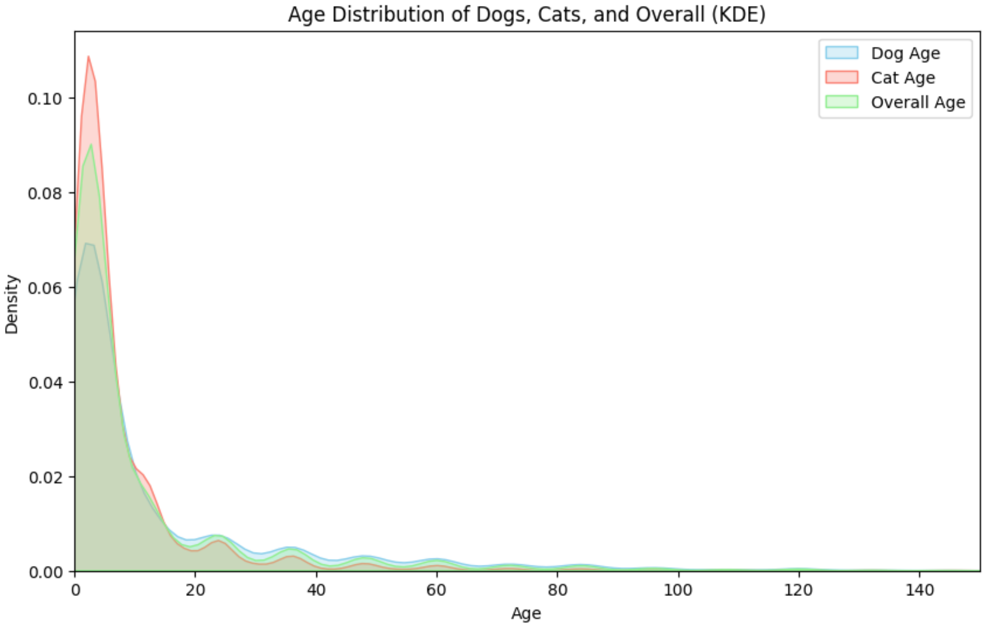


Figure 1: Age Distribution of Dogs, Cats, and Overall (KDE)

3.2 Machine Learning Models

3.2.1 Random Forest

In addition to traditional statistical analysis, we developed a multiclass Random Forest model to predict adoption speed, utilizing *multi:softmax* as the objective function and *mlogloss* as the evaluation metric. To optimize model performance, we performed a random search across 100 hyperparameter combinations, including *n_estimators*, *max_depth*, and *learning rate*. The model also incorporates optional sentiment-related features for added flexibility.

To gain deeper insights into feature interactions and their influence on predictions, we conducted a SHAP (SHapley Additive exPlanations) analysis. This analysis uncovered complex relationships between key attributes such as pet type, gender, age, and breed characteristics, highlighting their varying impacts on adoption speed.

3.2.2 GAM (Generalized Additive Models)

A Logistic Generalized Additive Model (GAM) was utilized to predict adoption speed. Given the dataset's high dimensionality—exceeding 300 features following one-hot encoding—we meticulously balanced feature complexity with model interpretability. For categorical variables unsuitable for ordinal encoding, such as color and breed, we applied target encoding to distill these attributes into single numerical values, thereby retaining their predictive relationship with the target variable.

To enhance model performance, we optimized the smoothing parameters through a grid search approach integrated within the GAM framework, achieving an effective equilibrium between flexibility and predictive accuracy. The inherent interpretability of GAM facilitated deeper analysis, including an assessment of the deviance explained summary and the computation of drop-in-accuracy metrics for individual features. To ensure coherence, related features—such as *Color1*, *Color2*, and *Color3*, which collectively characterize a pet's coloration—were aggregated in the drop-in-accuracy evaluation. Comprehensive results of this analysis will be elaborated in the subsequent section.

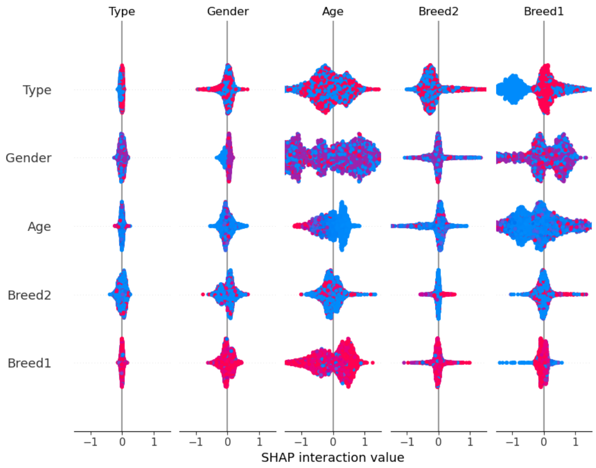
3.2.3 EBM (Explainable Boosting Machine)

The Explainable Boosting Machine (EBM) is a tree-based, cyclic gradient boosting model equipped with automatic interaction detection capabilities. Its tree-based design obviates the need for the elaborate encoding processes demanded by logistic Generalized Additive Models (GAMs), thereby facilitating a more straightforward implementation. Although we conducted experiments to fine-tune hyperparameters—such as maximum leaf count, smoothing rounds, and learning rate—the default settings consistently demonstrated optimal performance. Once trained, the EBM yields a lucid graphical representation of feature importance, which will be elaborated upon in the following section.

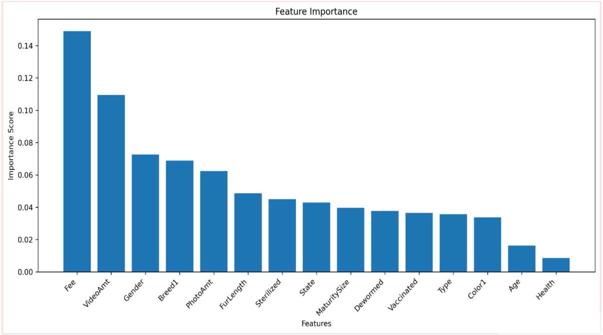
4 Narrative and Insights

In this section, we further analyze the results from the previous section and elucidate the underlying logic.

4.1 Random Forest



(a) SHAP Plots for Random Forest

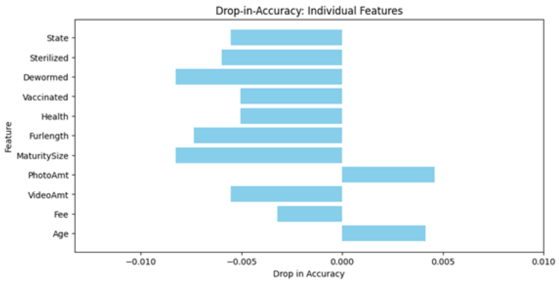


(b) Feature Importance for Random Forest

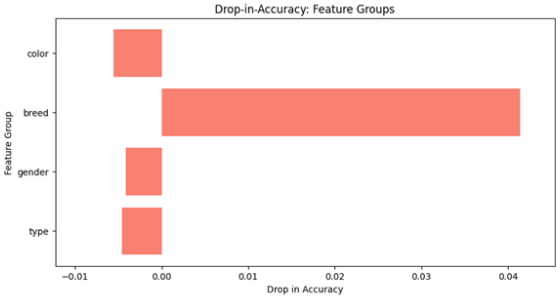
Feature importance analysis indicates that adoption fee emerges as the most influential predictor of pet adoption speed, followed closely by video content (*VideoAmt*) and gender. Breed characteristics and the number of photos demonstrate moderate predictive significance. Physical attributes, such as fur length and sterilization status, exert a lesser yet still discernible impact. Unexpectedly, health status and age rank as the least influential factors in determining adoption speed. To mitigate class imbalance within the adoption speed categories, we employed the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples for class 0, thereby ensuring a balanced dataset for model training.

SHAP (SHapley Additive exPlanations) analysis identifies *Type* as a pivotal predictor, exhibiting substantial interactions, particularly with *Breed1* and *Age*. These interactions underscore the model's capacity to effectively differentiate between pet categories. Gender-related interactions reveal balanced distributional patterns, suggesting equitable predictive behavior across gender classifications. *Age* displays intricate interaction dynamics, notably with *Type* and *Breed1*, highlighting subtle, category-specific variations influenced by age. Additionally, *Breed1* exhibits more pronounced interaction effects compared to *Breed2*, affirming that primary breed characteristics wield greater predictive influence than secondary ones.

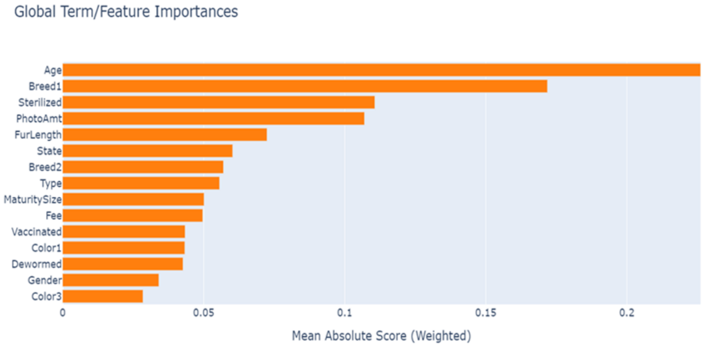
4.2 GAM and EBM



(a) GAM Feature Importance for Individual Features



(b) GAM Feature Importance for Grouped Features



(c) EBM Feature Importance

As illustrated in Figures 3a, 3b, and 3c, the Explainable Boosting Machine (EBM) identifies sterilization as a notable factor, yet both the Generalized Additive Model (GAM) and EBM consistently rank age, breed, and photo quantity within a pet's profile as the predominant features influencing adoption speed. This finding contrasts with the feature importance outcomes derived from the Random Forest model, a discrepancy likely attributable to data imbalance. As previously noted, the dataset is heavily skewed toward pets under 12 months of age. While the Random Forest model employed the Synthetic Minority Oversampling Technique (SMOTE) to address this imbalance, no equivalent adjustments were implemented during the training of GAM and EBM.

Notably, the results derived from the unbalanced data align with real-world patterns. Firstly, shelters frequently receive a disproportionate number of kittens and puppies, often due to stray mother cats or dogs producing multiple offspring that are subsequently relinquished. In some instances, these litters originate from owned animals whose caretakers are unable to sustain the additional burden. Secondly, younger pets tend to be adopted more rapidly, as their perceived appeal—enhanced by compelling photos or videos—captivates prospective adopters more effectively.

Despite variations in the emphasis placed on age, all models consistently highlight the critical role of multimedia in accentuating each pet's distinct beauty and personality. Consequently, shelters are encouraged to prioritize the development of engaging profiles enriched with high-quality photos and videos, as such efforts can significantly accelerate adoption rates.

5 Discussion and Reflections

Throughout this analysis, several significant challenges and limitations emerged. The foremost challenge involved mitigating data imbalance, particularly within age distributions and adoption speed categories, necessitating meticulous application of preprocessing techniques such as the Synthetic Minority Oversampling Technique (SMOTE). A further limitation stemmed from the geographic specificity of the dataset, which may not comprehensively reflect global adoption trends. Additionally, the exclusion of multi-pet profiles—while essential for maintaining analytical clarity—may have disregarded distinctive dynamics inherent to group adoptions.

Future investigations could enhance their scope by incorporating geographically diverse datasets and employing advanced methodologies to examine multi-pet adoption scenarios. Moreover, the integration of deep learning approaches for pet photo analysis, coupled with sophisticated natural language processing techniques for evaluating textual descriptions, could yield richer insights into the visual and narrative elements influencing adoption speed.

References

[1] Diesel, G., Pfeiffer, D. U., & Brodbelt, D. (2008). Factors affecting the success of rehoming dogs in the UK during 2005. *Preventive Veterinary Medicine*, 84(3-4), 228–241.

[2] Dinwoodie, I. R., Zottola, V., Kubitz, K., & Dodman, N. H. (2022). Selection factors influencing eventual owner satisfaction about pet dog adoption. *Animals*, 12(17), 2264.

[3] Gourkow, N. (2001). *Factors affecting the welfare and adoption rate of cats in an animal shelter* (Doctoral dissertation, University of British Columbia).

[4] Normando, S., Stefanini, C., Meers, L., Adamelli, S., Coultis, D., & Bono, G. (2006). Some factors influencing adoption of sheltered dogs. *Anthrozoös*, 19(3), 211–224.

[5] Zadeh, A., Combs, K., Burkey, B., Dop, J., Duffy, K., & Nosoudi, N. (2022). Pet analytics: Predicting adoption speed of pets from their online profiles. *Expert Systems with Applications*, 204, 117596.