Introduction

Every animal deserves a chance at a permanent home. Yet, year after year, shelters are filled beyond capacity, and far too many dogs and cats wait or never leave at all. Behind every kennel door is a life whose fate hinges on a handful of moments and decisions: Who is noticed, and who is passed by?

Our Research Question

Which intake-time characteristics of dogs and cats - age, sex, color, breed mix, health status, and intake condition most strongly influence their chances of being adopted? By quantifying the impact of each feature, we aim to predict an animal's adoption probability the day it arrives at the shelter. Drawing on his day-to-day experience, Ron S. from Adopt Me notes: *"From what we surface in the app, internal traits—age, size, basic training—seem to drive adopter clicks far more than timing or crowding."*

Why It Matters

Euthanasia of otherwise adoptable animals is a systemic issue. If shelters know in advance which traits lower the odds of adoption,they can intervene:

- Promote high-probability animals as "ambassadors" to draw visitors.
- Prioritize low-probability animals for enhanced visibility, enrichment, or targeted marketing.
- Data-driven triage can quite literally be the difference between life and death.

Why It's Hard

Real-world shelter data are messy and highly imbalanced:

- Class imbalance - far more "Adopted" cases than failures (or vice versa, depending on the shelter).
- Incomplete records - health notes, intake conditions, or breed labels can be missing or inconsistent.
- Unmeasured factors - charisma, visitor–pet chemistry, and staff advocacy are hard to encode.

What does theory / prior work tell us about this problem

Lepper et al. (2002) found age and color predicted dog adoption: "*In descending order of importance, the other predictors of adoption were youth and not having a primarily black coat*" (Lepper, Kass, & Hart, 2002). Carini et al. (2020) showed black cats had worse outcomes: "*Specifically, black cats experienced the highest euthanasia and lowest adoption rates, while white cats had the lowest euthanasia and highest adoption rates*" (Carini, Sinski, & Weber, 2020).

Our Approach

1. Curate key features (animal type, age, sex, color ,breed mix/purebred, health condition).
2. Handle imbalance with SMOTE inside each cross-validation fold to avoid data leakage.
3. Model with XGBoost-a tree-based ensemble robust to skewed predictor distributions.
4. Evaluate rigorously (10-fold CV, ROC-AUC, PR-AUC, confusion matrix).
5. Interpret results via feature importance and partial-dependence to provide actionable insights for shelter staff.

By focusing squarely on how specific traits raise or lower adoption odds, we move beyond descriptive statistics toward practical, ethically grounded intervention.

Data Overview

About The Data:

```
str(df)
```

```
## tibble [36,682 × 9] (S3: tbl_df/tbl/data.frame)
##  $ outcome_type  : Factor w/ 2 levels "Adopted","Not_Adopted": 1 2 1 2 2 2 2 1 2 1 ...
##   ..- attr(*, "names")= chr [1:36682] "Adoption" "Euthanasia" "Adoption" "Euthanasia" ...
##  $ animal_type   : Factor w/ 2 levels "Cat","Dog": 2 2 2 1 2 2 2 2 2 2 ...
##  $ mixed_purebred: Factor w/ 2 levels "Mixed","Purebred": 1 1 2 2 2 2 1 1 2 1 ...
##  $ color_group   : Factor w/ 3 levels "Black","Brown",..: 1 1 1 1 2 2 2 2 2 1 ...
##  $ health_status : Factor w/ 2 levels "Healthy","Sick_Aged_Pregnant": 1 2 2 2 2 2 2 2 1 2 1 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 2 2 2 1 2 2 1 2 2 1 ...
##  $ neutered      : Factor w/ 2 levels "Intact","Neutered_Spayed": 2 2 2 2 2 2 2 2 2 1 2 ...
##  $ age_norm      : num [1:36682] 0.727 0.636 0.773 0.864 0.682 ...
##  $ age_group     : Factor w/ 4 levels "Adult","Juvenile",..: 4 4 4 4 4 4 4 4 4 4 ...
##  - attr(*, "na.action")= 'omit' Named int [1:6] 6405 41824 55745 67133 67137 67139
##   ..- attr(*, "names")= chr [1:6] "6405" "41824" "55745" "67133" ...
```

Data Shape - 9 columns, 36,682 rows.

The primary entities in the data are animals, specifically focusing on Dogs and Cats.
Dogs: 22,394 entities
Cats: 14,630 entities

The dataset includes one target and 8 features, which can be grouped as follows:
Outcome Information: outcome_type (target)
Animal Characteristics: animal_type, mixed_purebred, color_group, sex, neutered, age_group, age_norm,
Health Status: health_status

```
summary(df)
```

```
##       outcome_type   animal_type  mixed_purebred  color_group
##  Adopted    :33434   Cat:14328   Mixed   :35176   Black:11593
##  Not_Adopted: 3248   Dog:22354   Purebred: 1506   Brown:18506
##                                                   Light: 6583
##
##
##
##             health_status       sex                   neutered
##  Healthy            :33056   Female:18056   Intact         :26611
##  Sick_Aged_Pregnant: 3626    Male  :18626   Neutered_Spayed:10071
##
##
##
##
##     age_norm              age_group
##  Min.   :0.000000   Adult       : 5643
##  1st Qu.:0.007472   Juvenile    :13145
##  Median :0.045455   Puppy_Kitten:15729
##  Mean   :0.082538   Senior      : 2165
##  3rd Qu.:0.090909
##  Max.   :1.000000
```

The data is clean, and has 0 Null values.

Most of the data is inherently imbalanced due to the typical characteristics of animals that end up in shelters.

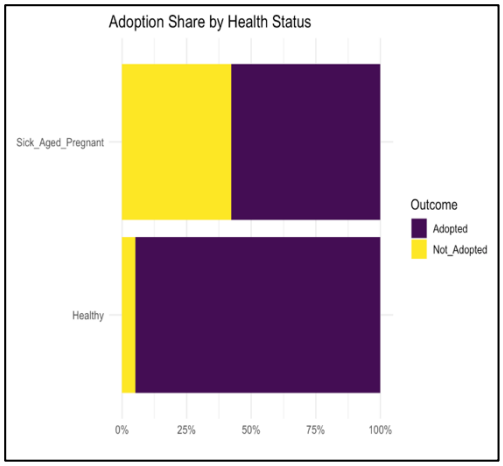
We address the class imbalance within the outcome_type target variable by incorporating SMOTE directly into our cross-validation training process.

For these imbalances within the explanatory features, we leverage the inherent robustness of the XGBoost algorithm. Tree-based ensemble methods like XGBoost are generally less sensitive to skewed or imbalanced distributions in predictor variables compared to some other machine learning algorithms.

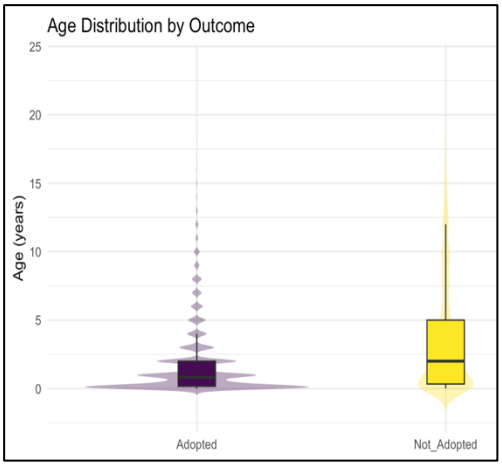
```
colSums(is.na(df))
```

```
##   outcome_type    animal_type mixed_purebred    color_group  health_status
##              0              0              0              0              0
##            sex       neutered       age_norm      age_group
##              0              0              0              0
```

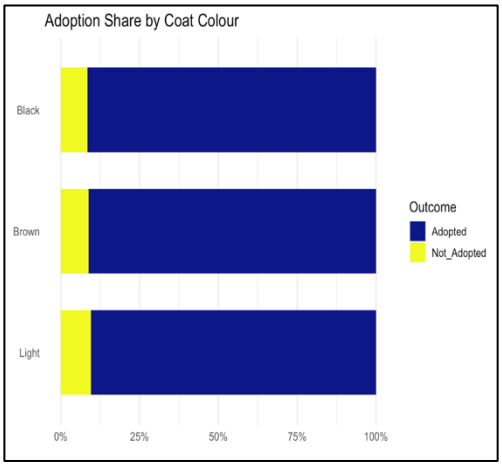
Data Visualization:



Healthy animals see adoption shares above 90 %, whereas the "Sick / Aged / Pregnant" group drops below 40 %, underscoring the clinical hurdle to placement.



Adopted animals are markedly younger on average than those not adopted, hinting that age will be a key predictor.



Across all major coat colours, adoption shares cluster tightly around 80 %, indicating that colour no longer materially affects placement likelihood.

Methods and Results

Modelling strategy:

To predict the final outcome type of each animal, a gradient-boosted tree model (XGBoost) was selected for its ability to capture nonlinear interactions and handle mixed feature types with minimal manual feature engineering.

Ten-fold cross-validation (stratified by outcome) on the 80 % training partition supplies out-of-sample estimates while a racing-ANOVA tuner explores 100+ XGBoost configurations, eliminating weak settings early. The search optimises ROC-AUC but tracks PR-AUC, accuracy, sensitivity, specificity, and F-measure to ensure balanced performance.

Best model configuration:

Parameter	Value	Notes
Trees	874	Enough depth for diminishing gains beyond 900
Max depth	7	Captures higher-order interactions without over-fitting
Learn rate (η)	0.046	Conservatively low for smoother convergence
mtry	58	≈ √(number of predictors) after dummy expansion
Min n (leaf size)	12	Prevents tiny, noisy leaves
γ loss-reduction	1.4	Prunes weak splits
Sample rate	0.80	Adds stochasticity, reducing variance

Model Results:

Metric	Score	Interpretation
Accuracy	**0.786**	≈ 4 of every 5 predictions are correct.
Sensitivity (Recall)	**0.796**	79.6 % of animals that were *actually adopted* were predicted correctly (5 333 / 6 700).
Specificity	**0.683**	68.3 % of *non-adopted* animals were flagged correctly (435 / 637).
F-measure	**0.872**	Strong balance between precision and recall.
ROC-AUC	**0.822**	Good class separation across all thresholds.
PR-AUC	**0.975**	Excellent precision even when recall is pushed high-crucial under heavy class imbalance.

The corresponding confusion matrix is:

	Truth = Adopted	**Truth = Not Adopted**
Pred = Adopted	**5 333** (TP)	202 (FP)
Pred = Not Adopted	1 367 (FN)	**435** (TN)

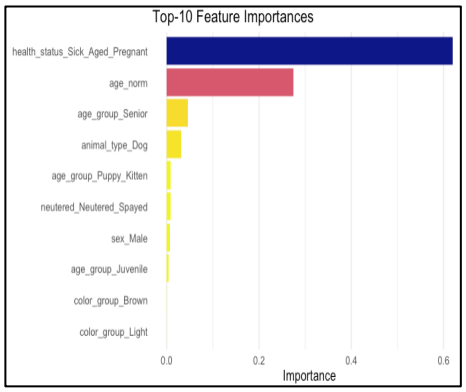
These counts confirm the sensitivity/specificity trade-off: the model deliberately tolerates some false positives (202) to avoid missing too many genuine adoptions.

Cross-validation and hold-out scores differ by less than one percentage point on every metric, indicating the model generalises well and is not over-tuned. Combined with the sky-high PR-AUC, these results suggest the model can be relied upon in an operational setting where the cost of missing an adoption opportunity outweighs a limited number of false alarms.

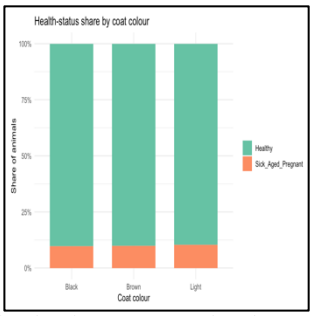
What drives the prediction?

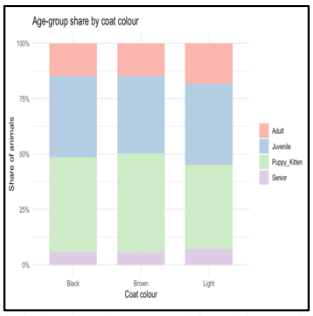
Rank	Feature	Relative Importance
1	health_status = Sick/Aged/Pregnant	~ 0.62
2	Age_norm	~ 0.26
3	age_group = Senior	~ 0.04
4	animal_type = Dog	~ 0.03
5+	All remaining predictors	< 0.01 each

The dominance of health_status_Sick_Aged_Pregnant aligns with shelter intuition-sick or frail animals face lower adoption odds-while the strong second-place showing of normalised age reiterates that younger animals are favoured. The long tail of near-zero features implies that, after accounting for health and age, additional gains come from only a handful of categorical signals (species, colour group, etc.). The long tail of near-zero features also reveals something counter-intuitive: coat colour contributes almost nothing to predictive power. This contradicts much of the earlier "black-dog bias" literature, which reported that dark-coloured (especially black) dogs are adopted less frequently. One plausible explanation is that public-awareness campaigns in recent years have reduced colour-based stigma, leading adopters to focus more on health and age. Alternatively, the effect may be shelter-specific-if intake procedures, lighting, or photo quality have improved, colour differences become less salient. Either way, the present results suggest that, in this dataset, coat colour is no longer a key determinant of adoption likelihood. Reacting to our findings, Ron S. (student volunteer, Adopt Me) admits: "*I always assumed coat colour mattered—maybe because people link it to breed—but the data say otherwise. Knowing that age and health status matter most lets us spotlight healthy middle-aged dogs that still struggle to find homes*."



The variable-importance plot shows a steep drop-off after the top two features





Animals of every coat color share virtually identical age and health profiles, confirming that color plays no meaningful role in adoption outcomes.

Limitations and Future Work

Limitations:

Our analysis is based on data collected from a single shelter in Austin, Texas. As such, the findings may not generalize well to other regions with different demographics, cultural attitudes toward pet adoption, or shelter practices. Additionally, external factors that may influence adoption rates such as local adoption campaigns, media events or even behavioral aspects like the dog's temperament, energy level, or the emotional "connection" perceived by potential adopters that are not captured in our dataset. These intangible and subjective elements are difficult to quantify and often go unrecorded, yet they may play a crucial role in real world adoption outcomes. Their absence may introduce hidden biases and limit the predictive power of our models.

Future Work:

Given additional time, we would aim to extend the dataset to include information from multiple animal shelters across diverse geographic regions worldwide. This would help minimize the influence of region specific factors such as local adoption campaigns and allow us to build a model that generalizes more effectively across different parts of the world. Ultimately, such a model would be better equipped to provide accurate predictions and actionable insights across a broader range of real-world contexts.