Dog + Pet Adoptions

Andrew Kim, Olivia Kris, Chae Won Yun

Problem

- Shelters are incredibly **overcrowded,** resulting in many dogs and other pets being euthanized due to lack of shelter space/lack of adoptions
- **Exploring to determine which attributes make prospective dog owners more inclined to adopt a dog**
    - Helps shelters **reduce likelihood of euthanization**
- *Key Stakeholders:* those running shelters, those seeking to adopt a dog

Why Machine Learning?

- Can help us better understand what qualities increase the likelihood of adoption
- Allows us to take a dataset with multiple factors (ex. Breed, size, etc.) to compile an accurate model of how each pet will do in the adoption process

Data Cleaning + First Steps

- Datasets were found on **Kaggle**
  - collected the information of adoptable dogs off of Petfinder.com's API
- Dataset provides columns describing attributes of dogs in the API.
  - Ex: breed, color, state where they were found, age, etc.
- Merged datasets on unique IDs of pets, called it all_dogs
- Filtered through duplicate ID values and removed them
- Discovered faulty data, such as values being in the wrong column, by using .value_counts and removed these values from our rows
- Filled any remaining nan values with "unknown" or False using the .fillna function

```python
# loading in dog description dataset

dog_desc = pd.read_csv('allDogDescriptions.csv')
dog_desc
```

Visualize

	index int64 0 - 58179	id int64	org_id object	url object	type.x object	species object	breed_primary obj...	b
140	140	45966526	NV184	https://www.petfi...	Dog	Dog	Labrador Retriever	r
141	141	45966461	NV184	https://www.petfi...	Dog	Dog	Terrier	M
142	142	45962677	NV26	https://www.petfi...	Dog	Dog	Mastiff	M
143	143	45962675	NV26	https://www.petfi...	Dog	Dog	Pit Bull Terrier	M
144	144	45961141	AZ189	https://www.petfi...	Dog	Dog	Labrador Retriever	r
145	145	45959652	NV202	https://www.petfi...	Dog	Dog	Redbone Coonho...	S
146	146	45957822	NV22	https://www.petfi...	Dog	Dog	Miniature Pinscher	r
147	147	45956950	AZ189	https://www.petfi...	Dog	Dog	Boxer	r
148	148	45956409	NV155	https://www.petfi...	Dog	Dog	Bichon Frise	r
149	149	45956384	NV155	https://www.petfi...	Dog	Dog	Pit Bull Terrier	r

58180 rows, showing 10 per page

Page 15 of 5818

Initial Data Prior to Cleaning

Data Cleaning: Filtering

- *Irrelevant*
  - Ex. ‘Type.x’ and ‘Type.y’ were all type ‘Dog,’ which remained constant
- *Inconsistencies across data*
  - Ex. ‘Breed_mixed’ was dropped due to ‘breed_primary’ and ‘breed_secondary,’ as we saw discrepancies between the relationship between the two where ‘breed_mixed’ was set to False when there were secondary breeds involved
- *Only nan values*
  - Ex. ‘Declawed’ had only nan values, which was not useful to our ML process
- *Too many unique values that were not relevant*
  - Ex. ‘Found’ had the specific street name and situation in which the dogs were found, but there are too many unique values and this is not relevant in someone’s decision to adopt a dog

```python
# filling nan values with "unknown" or False


values = {"env_children": False, "env_dogs": False, "env_cats": False, 'remove': False}
all_dogs = all_dogs.fillna(value=values)
all_dogs
```

Visualize

	id int64	contact_state_x ob...	description_x object	remove bool	breed_primary obj...	breed_secondary o...	color_primary object
	8619716 - 46043149	VA 11.6% NY 10.8% 42 others 77.6%	To adopt on... 0.4% Please call t... 0.3% 3921 others 99.3%	False 77.9% True 22.1%	Labrador R... 16.4% Chihuahua 7.5% 151 others 76.1%	Labrador Ret... 5.3% 115 others 32.2% Missing 62.5%	Black 14.4% 14 others 32.2% Missing 53.3%
529	45908253	GA	Thank you for yo...	True	Chihuahua	nan	nan
530	45908161	GA	Pepper is pure pe...	True	Mixed Breed	nan	Harlequin
531	45908093	GA	MUST BE ADOPT...	True	Cavalier King Cha...	Poodle	Yellow / Tan / Blo...
532	45907998	GA	MUST BE ADOPT...	True	Cavalier King Cha...	Bichon Frise	Apricot / Beige
533	45907992	GA	Hi! I am Elle and I ...	True	Great Dane	nan	Black
534	45907975	GA	Thank you for yo...	True	Silky Terrier	nan	nan
535	45907823	GA	Milly is a very pla...	True	Catahoula Leopar...	nan	Brown / Chocolate
536	45907805	GA	Hi, Iâ€ m Mia! I...	True	Scottish Terrier	nan	Black
537	45907698	GA	Meet Ripley! If yo...	True	Labrador Retriever	nan	Brindle
538	45830513	GA	Meet Speckles!S...	False	Spitz	Jindo	nan

4112 rows, showing 10 per page   Page 44 of 412

Post Data Cleaning

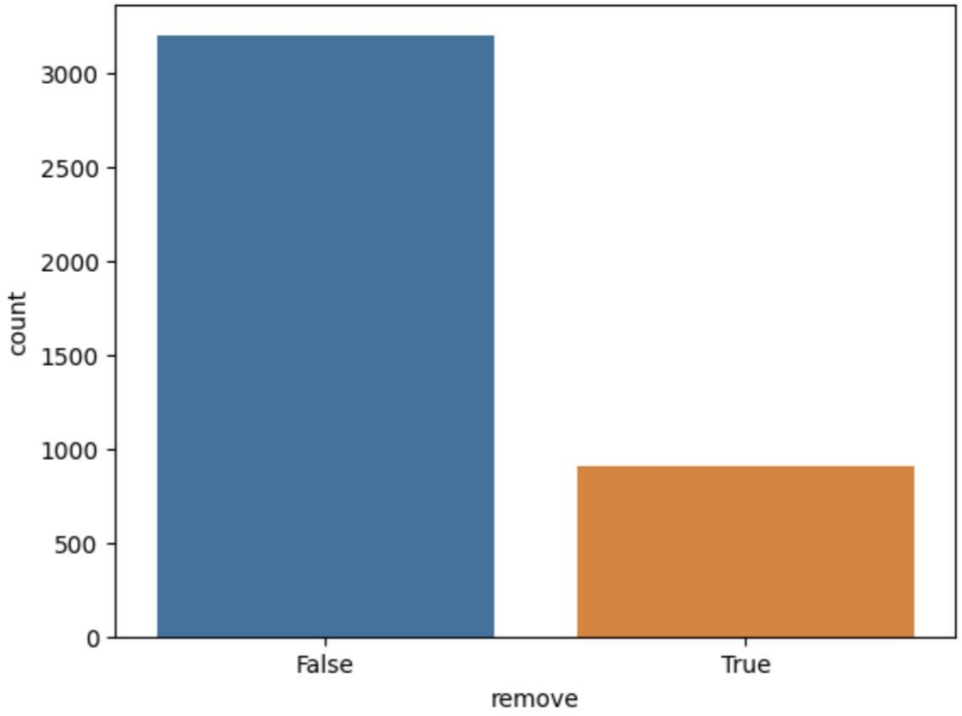
Methods

We want our model to be able to handle mostly categorical data such as breed type and color of dog

1. ***Baseline Model***
    a. the accuracy of predicting whether a dog will be adopted
    b. serves as a guideline for how powerful our models predictions are
2. ***Logistic Regression***
    a. Best suited for prediction and categorization problems
    b. ROC/AUC
3. ***CART/Decision Trees***
    a. Easily handles both categorical and continuous variables
    b. Ideal for larger datasets
4. ***Decision Trees with Cross-Validation***
    a. Gives us a better estimate of our trained model when applied to different datasets
    b. Protection against overfitting
    c. Pruning decision tree with ccp_alpha

Train-Test Splitting

1. Decided our y variable would be if a dog was removed or not from the shelter, meaning ***it was or was not adopted***
   a. Used ‘remove’ column
2. **Got dummy values of X**
   a. Used train_test_split to separate everything into X_train, X_test, y_train, y_test
3. **Split ⅓ of data to be used for validation**




```
#count of removed and not removed dogs

sns.countplot(x = y)
plt.show()
```

Visualization of Removed Dogs

```python
# baseline model
print(y_test.value_counts())
notremoved = np.sum(y_test == False)
removed = np.sum(y_test == True)

baseline_acc = notremoved / (notremoved + removed)

print(f'Baseline Accuracy: {baseline_acc}')
```

```
False    1039
True      318
Name: remove, dtype: int64
Baseline Accuracy: 0.765659543109801
```

Baseline Model Accuracy

```python

# logistic regression

logreg_model = LogisticRegression(random_state=42)
result = logreg_model.fit(X_train, y_train)

#logit_model=sm.Logit(y,X.astype(float))
#result=logit_model.fit()
#print(result.summary())

#logreg_model = smf.logit(formula = "remove ~ fixed + house_trained + \
#special_needs + shots_current", data = logmodeltemp).fit()

# make predictions
y_pred = logreg_model.predict(X_test)
#binary = [1 if x>= 0.5 else 0 for x in predictions]
# calculate accuracy
log_acc = accuracy_score(y_test, y_pred)

print(f'LogReg Test Accuracy: {log_acc}')

```

```
LogReg Test Accuracy: 0.7693441414885778
     Baseline Accuracy: 0.765659543109801
```

Logistic Regression Accuracy

```python
# baseline model
print(y_test.value_counts())
notremoved = np.sum(y_test == False)
removed = np.sum(y_test == True)

baseline_acc = notremoved / (notremoved + removed)

print(f'Baseline Accuracy: {baseline_acc}')
```

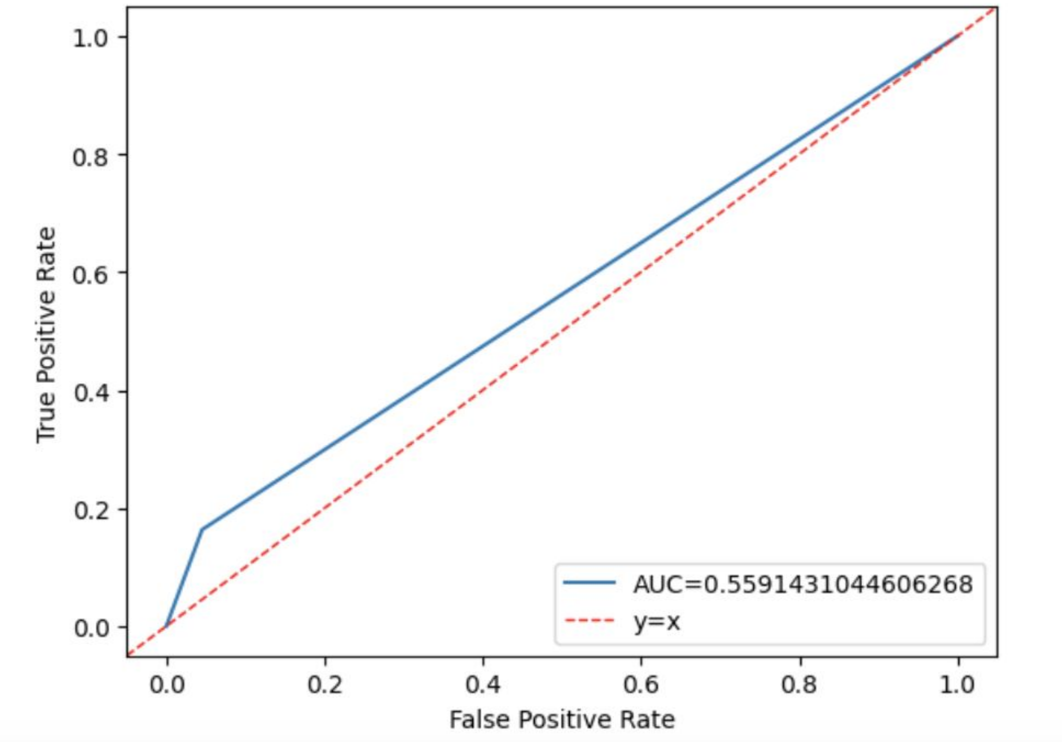
```
False     1039
True       318
Name: remove, dtype: int64
Baseline Accuracy: 0.765659543109801

LogReg Test Accuracy: 0.7693441414885778
```

```
LogReg Confusion Matrix:
 [[992  47]
 [266  52]]
True Positive Rate: 0.16352201257861634
False Positive Rate: 0.04523580365736285
True Negative Rate: 0.9547641963426372
False Negative Rate: 0.04523580365736285
```

Here we understand the reason behind our logistic regression model **not showing significant improvement over the baseline**

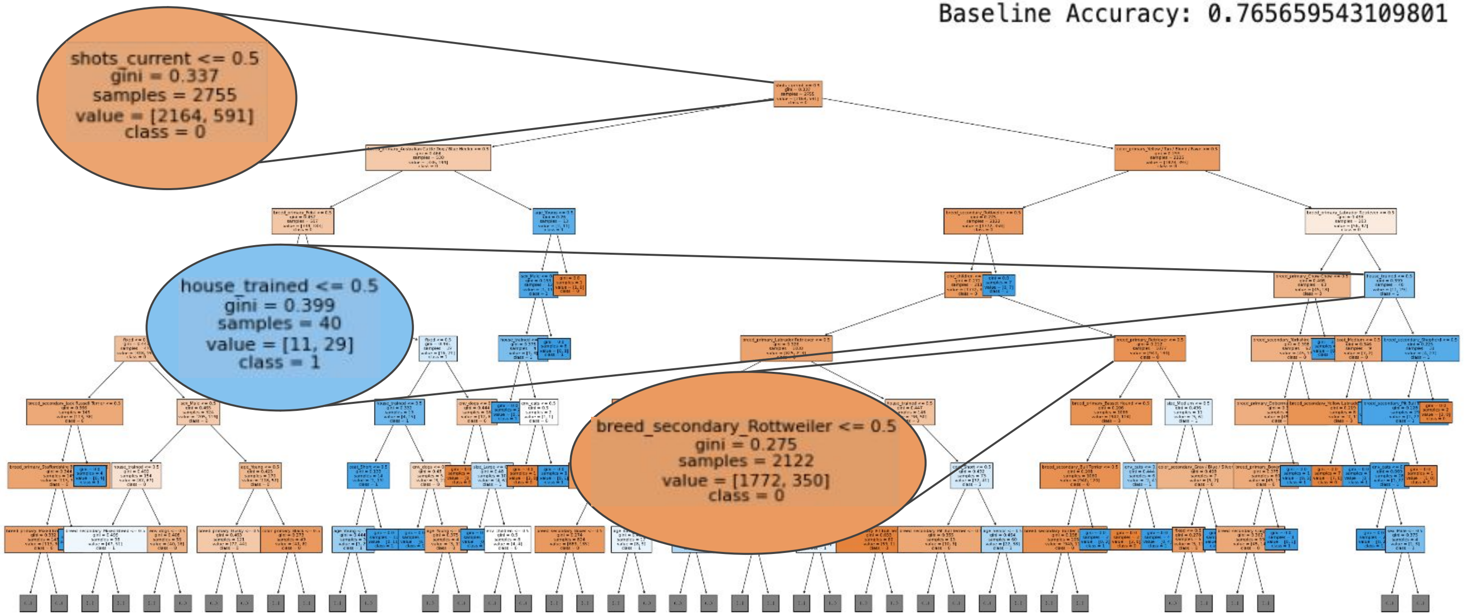
Baseline Performance, LogReg, and Confusion Matrix



Visualization of ROC Curve

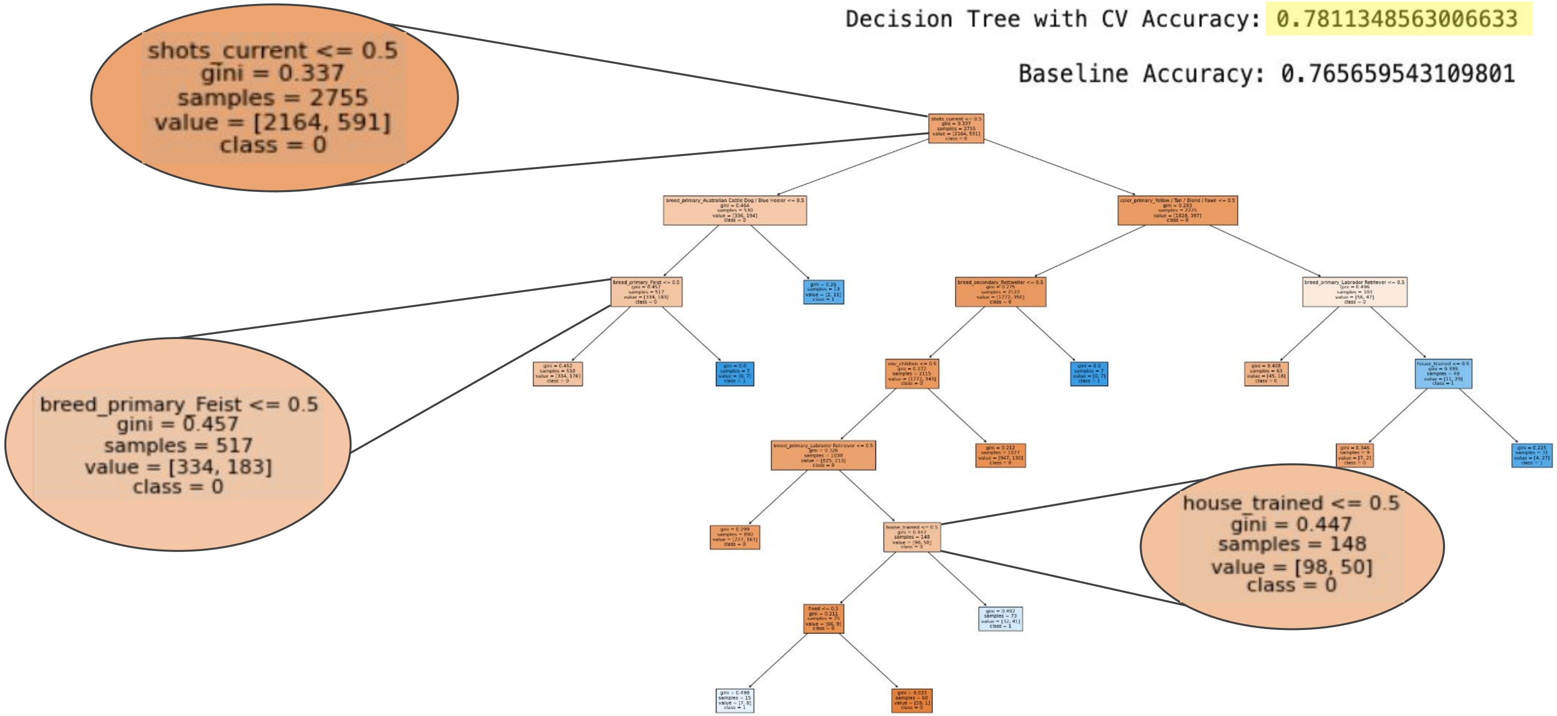
Decision Tree Accuracy: 0.7715549005158437

Baseline Accuracy: 0.765659543109801



Visualization of CART Decision Tree

Baseline Accuracy: 0.765659543109801



Visualization of Decision Tree with Cross Validation

Results

Baseline Model	Logistic Regression	CART	Cross Validated CART
0.765659543109801	0.7693441414885780	0.7715549005158440	0.7811348563006630

After testing all of our models, we concluded that the cross-validated CART model performed the best among the other models.

Results Cont.

Certain breeds of dogs are more likely to be adopted

- Due to the perceived image of a certain breed (ex: Golden Retrievers vs. Pitbulls)

Color of dog is likely to be highly correlated with the breed of dog

- one of our splits being color yellow/tan/blonde, which are primary colors of common friendly dog breeds such as golden retrievers (the split immediately following this is whether the dog is a labrador retriever)

If the dog is kept up on shots, house trained, good with children, and/or NOT fixed, they are more likely to be adopted

- It is much easier to go through with a spaying/neutering procedure in the future than to upkeep all missing shots, house train, and determine whether a dog is good around children

Conclusions + Implications

Implications of our models demonstrate that the breed of dog, whether they are **up to date on shots**, if they are **house trained**, if they are **good with children**, and if they are **fixed** are all strong determinants of whether a dog will be adopted or not.

Because of our findings from our models, our results are nearly ready to be used in the real world.

Results incentivise shelters to maintain their upkeep with shots, housetraining, etc

→

Increase the likelihood of Adoptable Dogs + Decrease the Number of Returned Shelter Dogs

Room for Improvement

- Addition of **sentimentality score** using description column (given to each dog) as another predictor
  - Out of scope
- Improved decision tree parameters
- We could have more definitive results if not for the fact that **values are occasionally missing for dogs**
  - **No even distribution** of the breeds of dogs available in our data
    - Ex: Labradors have 675 values in data compared to Bernese Mountain Dog with 1
  - Would be fixed with more data and consistently recorded data

Sources

Chauhan, A. (2022, October 28). *Dog adoption*. Kaggle. https://www.kaggle.com/datasets/whenamancodes/dog-adoption?select=dogTravel.csv

Databender. (2020, August 19). *PET adoption modified data cleaned*. Kaggle. https://www.kaggle.com/datasets/umairnsr87/petadoption-mod/data

Pet statistics. ASPCA. (n.d.). https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics

Thank you!

Any questions or comments?