PetFinder Challenge: Predicting Pet Adoption Speed

Sherine Zhang, Kaylee Zhang

{ sherinez, kayleez}@stanford.edu, Stanford University

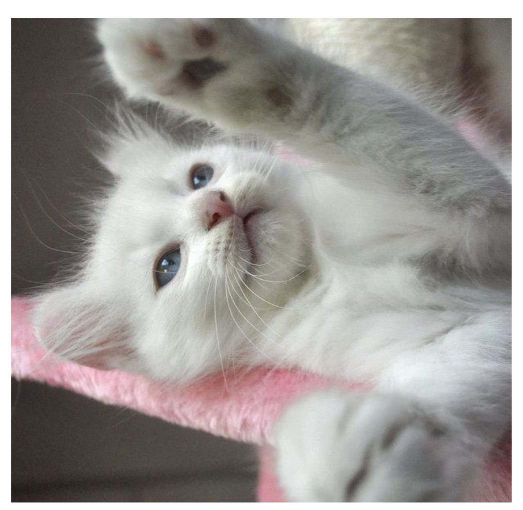
Problem & Task

PROBLEM: Predicting Pet Adoption Speed

- Our goal is to improve the adoption speed of stray animals by analyzing factors that affect it.
- Used traditional machine learning and deep learning methods
- Inputs: categorical data and text data
- Outputs: class labels in between 0 to 4

RESULTS

- Random forest and FC model perform the best among all machine learning and deep learning models respectively.



Features

CATEGORICAL FEATURES

- include animal type, breed, gender, color, maturity size, fur length, vaccinated, dewormed and more.
- We convert categorical data to one hot encodings in order to feed them into Deep Learning models and make Neural Network feasible.

TEXT FEATURES: description

- Used GloVe pre-trained model to produce word embeddings for the words in description.
- Each word is embedded to a size of 50.

Dataset

DATASET: Provided by PetFinder and Kaggle

- A csv file containing detailed information about the animals
- A json file containing descriptions with sentiment scores
- A large collection of videos and images of the animals

RESPONSE VARIABLE: Adoption Speed

- Divided into 5 categories: same day, within one week, 8 to 30 days, 30 to 90 days, no adoption after 100 days.

Models

TRADITIONAL MACHINE LEARNING MODELS

- Logistic regression

$$\Pr(Y_i = c|\mathbf{X};\beta) = \frac{e^{\beta_c \cdot \mathbf{X}_i}}{\sum_{k=1}^{K} e^{\beta_k \cdot \mathbf{X}_i}}$$

- Naive bayes and Support vector machines
- Decision trees:
  - Each split: reduce error & improve purity
  - Output variable importance
- Random forest and Gradient Boosting
  - Ensemble methods
  - Has smaller variance than Decision tree

DEEP LEARNING MODELS

- Fully Connected model with one-hot encodings
- LSTM with word embeddings from description
- Combined model

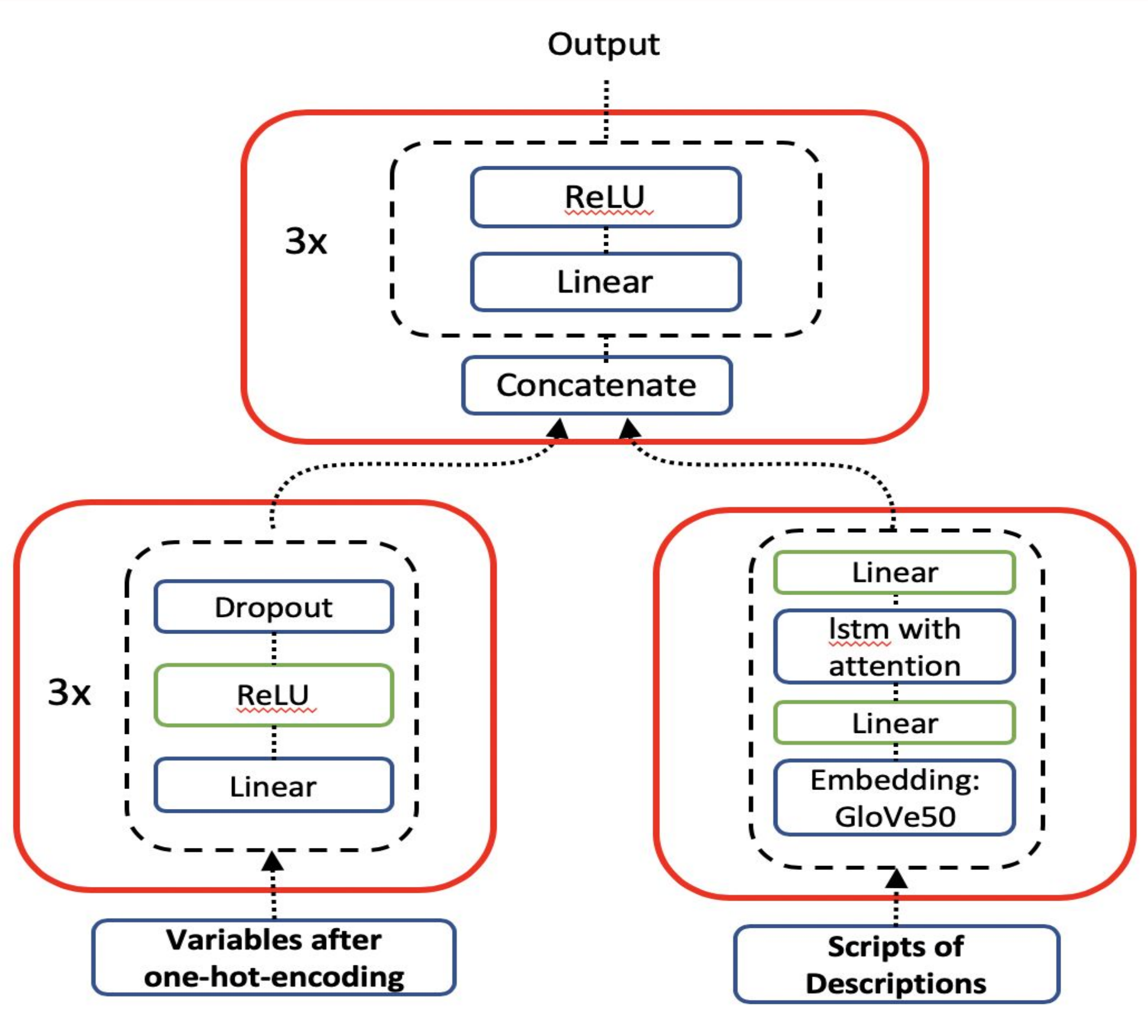


FIGURE 1: combined model pipeline. Fully Connected model on the bottom left, LSTM model on the bottom right.

Results

		Statistics		
		Precision	F1-score	Accuracy
Machine Learning Models	Logistic Regression	0.32	0.31	0.335
	Naïve Bayes	0.33	0.29	0.311
	SVM	0.34	0.33	0.359
	Decision Tree	0.33	0.32	0.321
	Random Forest	0.38	0.38	0.392
	Gradient Boosting	0.37	0.36	0.385
		Cross Entropy Loss	Evaluation Accuracy	Test Accuracy
Deep Learning Models	Fully Connected	0.00101	0.393	0.396
	LSTM	0.00097	0.294	0.322
	Combined	0.00091	0.383	0.384

TABLE 1: In total there are 14993 observations in the entire dataset. After splitting, 10795 examples are in the training set (72%), 2699 examples are in the validation set (18%), and 1499 examples are in the test set (10%).

Discussion

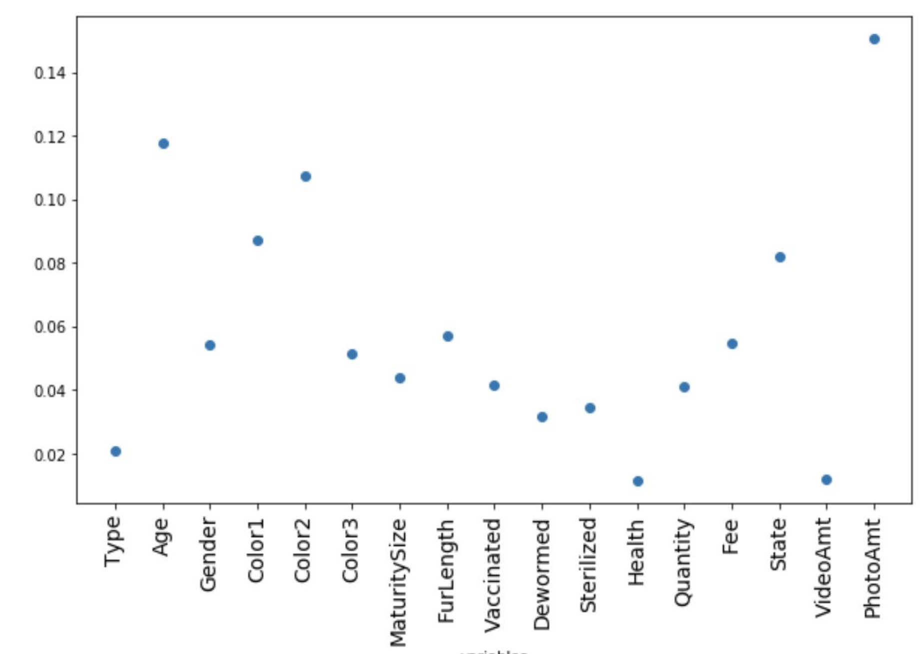


FIGURE 2: feature importance of decision tree model. It is calculated by the weighted decrease in node purity. Age, # photos are importance features.

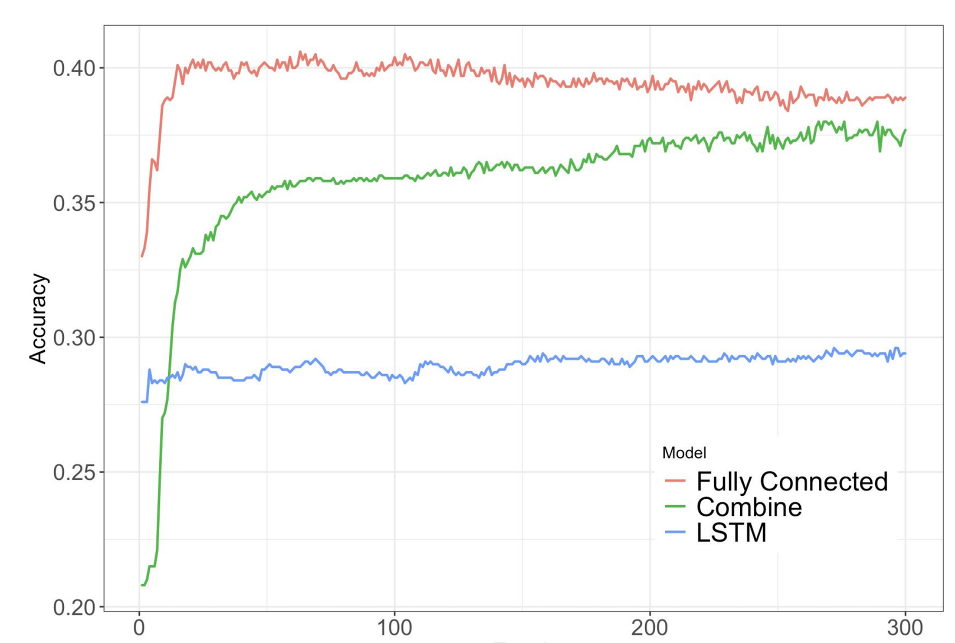


FIGURE 3: The trend of eval accuracy as epoch increases. The Fully Connected model converges the fastest. LSTM model on description script does not have learning trend.

	PREDICTED				
TRUE	0	28	7	3	20
	1	297	91	40	99
	1	363	133	80	160
	1	260	93	88	155
	1	288	104	64	322

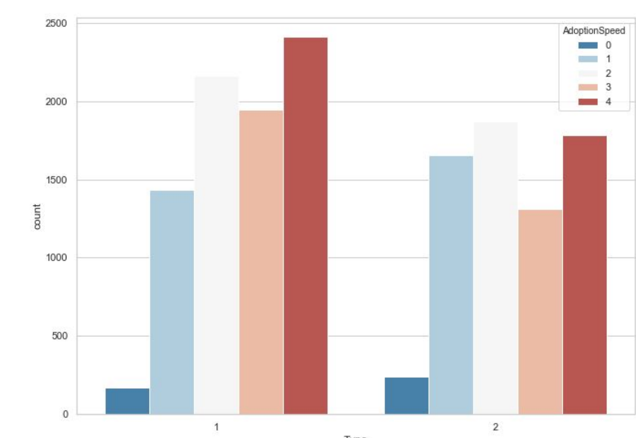


FIGURE 4: Left is the confusion matrix of Logistic Regression. Right is the distribution of 5 classes. Class 0 examples occupy a small proportion of total dataset, thus under 0/1 loss, the model classifies these examples to other classes to achieve better overall accuracy.

Future Work

- We would like to first experiment with more machine learning models and try different ensemble methods for a boost in performance.
- Given more time, we would be able to take full advantage of the available dataset and incorporate image and video data into our deep learning models with more hyperparameter tuning.

References

[1] Leo Breiman.Classification and regression trees. Routledge, 2017.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001.