Capstone Project 2
Petfinder Milestone Report 2

Table of Contents

Capstone Project 2
Petfinder Milestone Report 2

Objectives:

The objective for this report is to share machine learning results to build a model to predict the speed at which a pet is adopted. The classification areas are...

0 - Pet was adopted on the same day it was listed

1 - Pet was adopted between 1 and 7 days (1st week) after being listed

2 - Pet was adopted between 8 and 30 days (1st month) after being listed

3 - Pet was adopted between 31 and 90 days (2nd & 3rd month) after being listed

4 – No adoption after 100 days of being listed. (There are no pets in this dataset that waited between 90 and 100 days

About the Data:

Petfinder has provided the data at this link, which consists of 14,993 pets (54% dogs and 46% cats) that were up for adoption. Along with various features, this data also contains a column called "Adoption Speed", which is the predictor variable for this project.

Machine Learning

Models

The algorithms in the table below did not perform well when using almost all (20 features) of the training data with default parameters. Possible reasons for this poor performance are too many columns used and hyper-parameter tuning need to be applied.

Algorithm	Accuracy Score	Precision Score	Recall
Logistic Regression	35%	33%	35%
KNN	31%	34%	31%
Naive Bayes	34%	36%	34%
Random Forest	36%	35%	36%

Natural Language Processing (NLP) – Word2Vec

using Word2Vec a sentiment analysis was completed and identified word similarities such as the following...

"dog and cat" are 62% similar across the Description feature

"dog and pet" and "cat and pet" are 53% and 42% similar respectively.

Natural Language Processing (NLP) – tdidf vectorizer with a Naive-Bayes

the results of creating a vectorizer and using it in a Naive-Bayes algorithm yielded accuracy and confusion matrix results very similar to the prior models above ...

Capstone Project 2
Petfinder Milestone Report 2

Algorithm	Accuracy Score	Confusion Matrix
tfidf vectorizer with a Naive-Bayes algorithm	36%	*Adoption Speed* 0 1 2 3 4 0 **5** 33 33 19 41 1 4 **257** 348 129 280 2 2 220 **516** 260 392 3 0 126 321 **301** 292 4 3 177 301 185 **699**

Summary of Key Findings:

1. The algorithms (Logistic Regression, KNN, Naive-Bayes, and Random Forest) using 20 features yielded consist results in the range of 31 to 36% accuracy.
2. Natural Language Processing on the Description feature via creating a tfidf vectorizer and Naive Bayes yielded a similar accuracy result of 36%.
3. Needless to say, the above results are very disappointing and the next step is to identify what steps to take to try and improve performance. Ideas being considered are...
    a. hyper-parameter tuning
    b. reducing the number of features used
    c. combining the NLP Naive-Bayes with the other other algorithms in an ensemble
    d. moving on to deep learning (neural networks)

Links to Jupyter Machine Learning Notebooks (Python)

link folder for Milestone Report 2