PetFinder Adoption Prediction

How cute is that doggy
in the shelter?

AGENDA

1 Problem

2 Data

3 Exploring the Data

4 Prediction Models

5 Summary

6 Recommendations

Problem

Sponsor:

click to go to site

PetFinder is Malaysia’s leading animal welfare platform and collaborates closely with individuals and organizations to improve animal welfare.

Problem:

- This Kaggle competition is to develop algorithms to predict how quickly a pet is adopted
- The classification areas for adoption 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

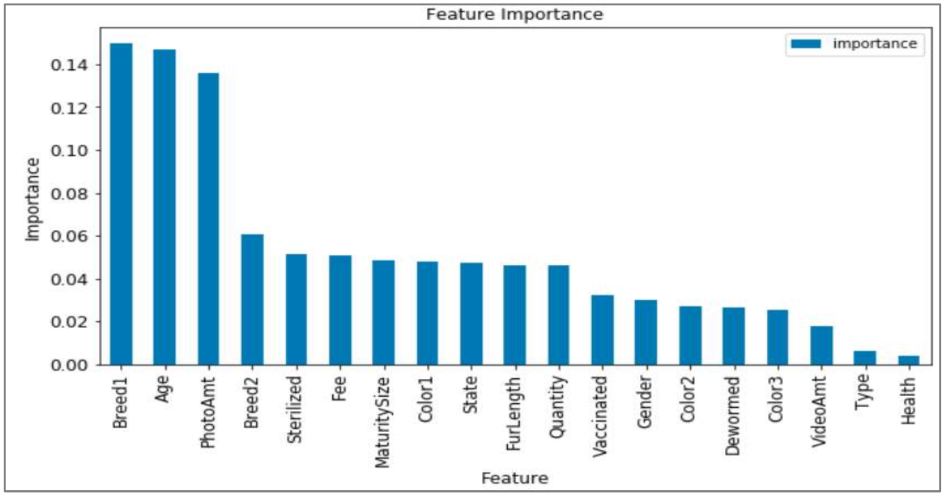
Quantity:

- 14,993 pets (54% dogs and 46% cats)

Data - Features

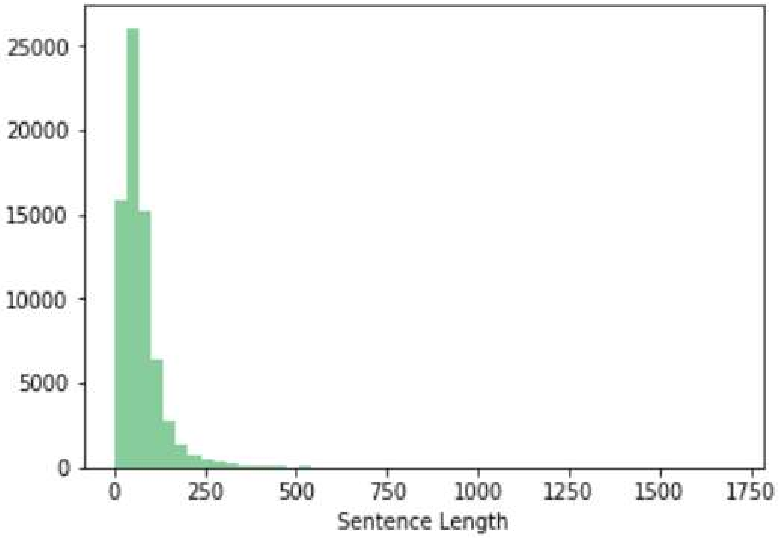
Non-Natural Language Processing

Feature importance breaks into three groups: Top 3, Medium 8, Low 8



Natural Language Processing

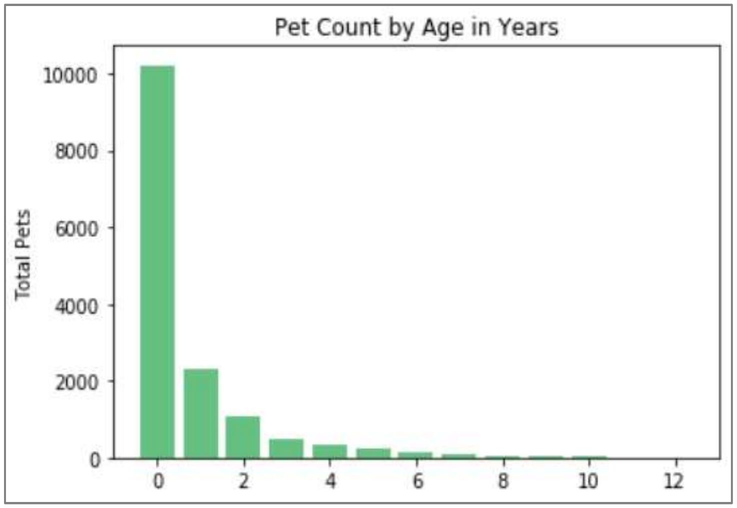
Descriptions is almost 70,000 sentences & over 900,000 words.



Predictor: Adoption Speed

Features not used: Pet id, Name, Rescuer Id

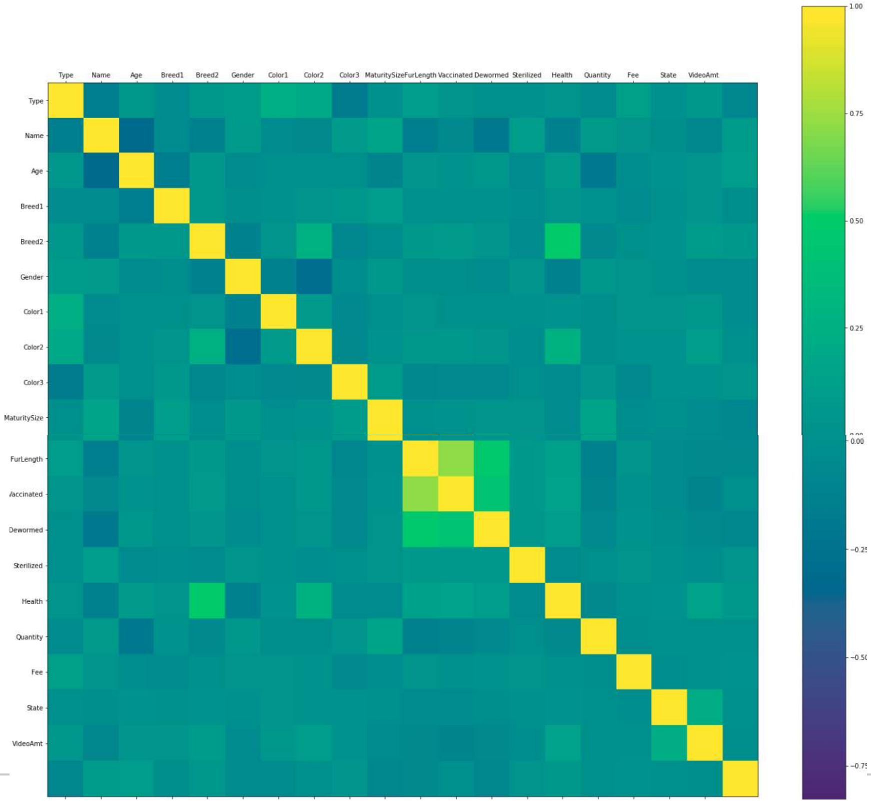
Exploring the Data – Top 3 Features

Breed	Age	Number of Photos
Need to add “Total pets by breed”		Need to add “Total cats & dogs with and without photos

Note: *For more information on these features, and all the other features refer to Milestone report 1.*

Exploring the Data – Correlation

- Though it's hard to read the feature names, this matrix illustrate that overall there is low correlation between the features, including the adoption speed.



Exploring the Data – Adoption (variable to predict)

Mean:

- The mean adoption speed is similar for dogs and cats at 2.6 and 2.4 (8-30 days to adopt)
- A comparison of the mean adoption speed by breed for dogs and cats shows a negligible difference in adoption speed. There is not a particular breed that deviates significantly from the mean.

Age:

- The ratio of adoption speeds for dogs vs cats for each year of age varies, but is not consistently higher or lower for one versus the other.
- Males of a given age have a better chance of being adopted than females (for some ages as high as 10% better)
- Across all ages (0 – 12 years) of pets, the number of medium size pets is always higher than the other types. As a result, the %NotAdopted pets is higher for medium dogs in over half the years.

Size:

- Mean adoption speed is similar across sizes, ranging from 2.35 to 2.57; which is also very close to the overall mean adoption speed for pets.

Media:

- Pets with no media (pictures nor video) have a mean adoption rate of 3.1 (31 to 90 days to adopt) versus the 2.5 (8 to 30 days to adopt) mean adoption rate across all pets.

Prediction Models – Round 1

Default Parameters:

- Given the low correlation (<12%) across features, this study will lean towards ensemble algorithms; premise being that weak individual predictors when combined will result in a predictor model of higher accuracy.

- The algorithms below did not perform well using 19 features and default parameters. As expected, the ensemble algorithm (Random Forest) outperformed the other algorithms, but only by a few percentage points.

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

Prediction Models – Round 2

Hyper-parameter Tuning for non-NLP and **Using Word2Vec with XGBoost for NLP:**

- Logistic Regression, KNN, & Naïve-Bayes were removed & XGBoost was added. This enabled a focus on ensembles.
- GridSearchCV was employed to find the best parameters for XGBoost and Random Forest.
- The hyper-parameter tuning did not change the results of accuracy scores for XGBoost nor for Random Forest.
- XGBoost with Word2Vec for NLP produced results similar to those of the Round 1 algorithms
- Introducing XGBoost provided a new “best model”, increasing accuracy by 4% and precision by 6%.

Algorithm	Accuracy Score	Precision Score	Recall
XGBoost	**40%**	**41%**	**40%**
Random Forest	36%	35%	36%
Word2Vec with XGBoost for NLP	35%	37%	35%

Note: *For more information, such as parameters tuned and values used to tune, refer to Milestone report 3.*

Prediction Models – Round 3

Stacking:

- The XGBoost model using 20 nonNLP features prediction results and the XGBoost model using NLP's prediction results were combined into an array, then used as input to a new XGBoost model.
- The results obtained were the best to date at 44% accuracy; a 4% improvement over the previous best model.

Algorithm	Accuracy Score	Precision Score	Recall
Stacked model	**44%**	**36%**	**44%**

Summary

Best model:

- The best model was the Scaled model that combined prediction results of the nonNLP XGBoost and NLP XGBoost models. Accuracy was 44%.

Algorithms:

- Using XGBoost on the nonNLP features paid off; it generated the best performing model using the 19 features (accuracy score of 40%) and is the only model to achieve 100% precision score in any adoption category (pets adopted on same day).
- Using NLP with Word2Vec & XGBoost on the Pet Description feature yielded a model that produced results in the mid-30's for accuracy, precision, and recall. Not a significantly different result from the nonNLP models.
- Hyper-parameter tuning did not significantly change the ensemble algorithms results.

Other:

- Removal of outliers did not significantly change the model results. This is probably due to the low number of only 15 outliers (pets over the age of 12) out of the 14,993 records.
- Adding Scalarization to the data did not significantly change the model results. This is probably due to most of the features are already in a tight range. Only the age of the pet was a range of values (0 to 12 years)

Recommendations

Algorithms:

- This analysis included 19 features to predict the adoption rate. Based on the feature importance results a fourth round using XGBoost and RandomForest with only the top three features; and possibly a fifth round using the top 11 features might improve the model's accuracy and precision.
- The best results were obtained by Scaling two models (one NLP and one nonNLP). Considering the breakout of Feature Importance into two primary contribution levels (features 1 – 3 in one group and features 4 – 11 in a second group), consider stacking 3 separate models – model for features 1 – 3, model for features 4 – 11, and the NLP with Word2Vec model.

NLP:

- Additional research into the contribution of Natural Language Processing (NLP) to model accuracy should be considered, specifically engagement of deep learning.

KNN:

- Consider running the KNN algorithm using only the top three features. The reduction in dimensionality may improve this algorithm's ability to group similar adoption trends, and produce a model with improved accuracy and precision.