Pet Adoption

A Data Science Capstone Presentation by:
Yuwen Dai, Allison Yuan, and Fred Watts

Quick Guess

- Maturity Size: 4
- Energetic Level: Needs a Lot of Activity
- Vaccinated: Yes
- Age: 4 months
- Dewormed: Yes
- Healthy Status: Yes
- Gender: Male

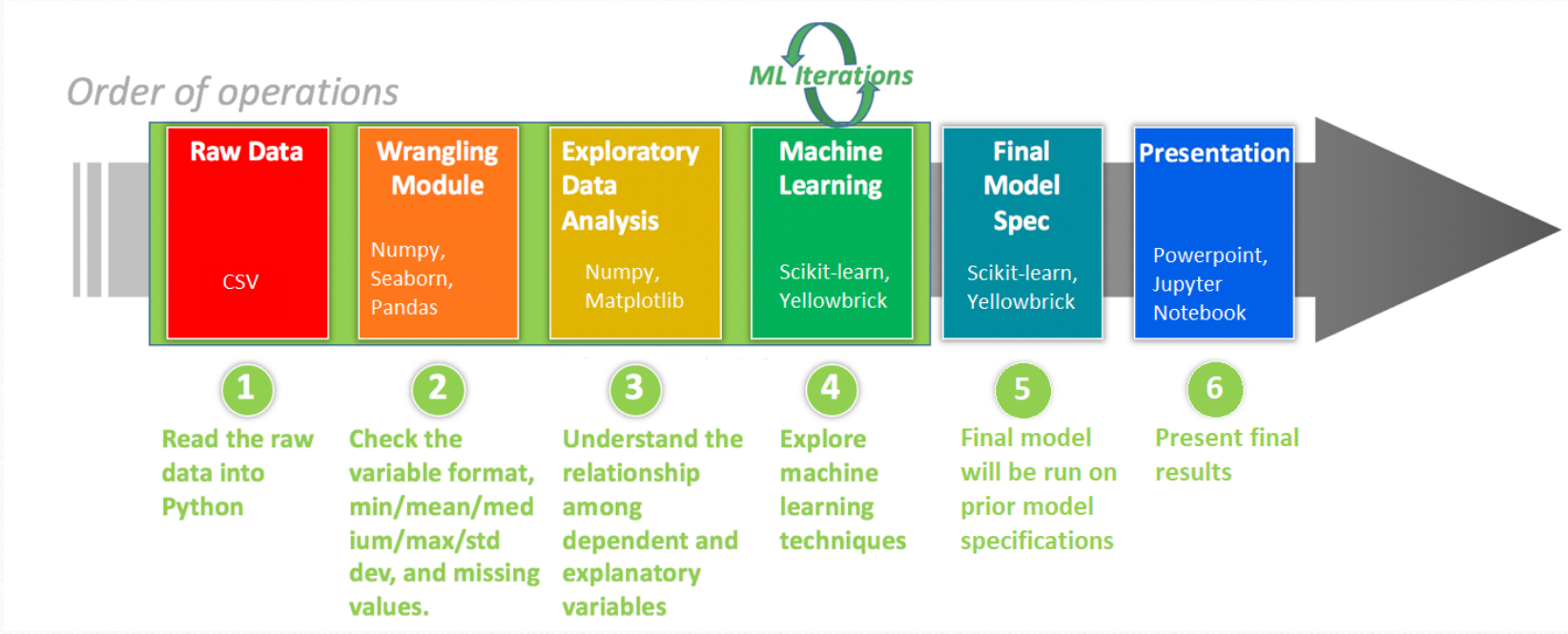
The Problem

- Domain: Animal Welfare
- Animal shelters are miserably overcrowded; many euthanize
- Shelters advertise pets using web profiles that include many traits (e.g., breed, age, vaccinations, etc.) as well as photographs and descriptions
- However, there is little evidence that profiles are designed with care for actual user preferences (i.e., they are uncurated info-dumps)

Hypothesis

- We believe that data science tools could identify which traits in a profile actually convince readers to adopt pets and which are ignored
- Animal shelters could build profiles to emphasize these key traits so pets will be adopted more quickly
- More adopted pets equals less misery and death

Pipeline Architecture



Raw Data Sources

- A dataset of Malaysian pet adoption profiles from Kaggle
  - ~15,000 pet profiles (exclusively cats and dogs)
  - 27 profile variables
- Datasets on cat and dog personalities by breed from the American Kennel Club and Cat Fanciers Association
- A dataset on Malaysian state populations from a Malaysian government census

Exploratory Data Analysis

- Our target variable was Adoption Speed

  0 - Pet was adopted on the same day as 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).

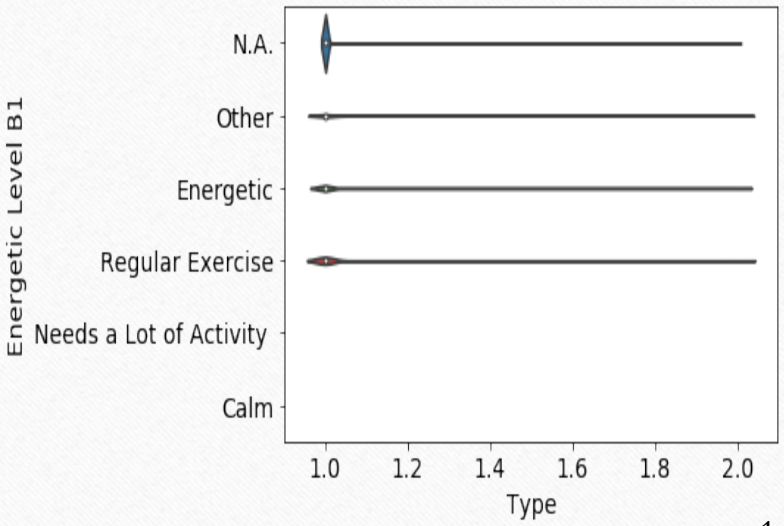
- Most of the explanatory variables are categorical variables, we first ran frequency tables on those variables to check count, row/column percentages by themselves, and cross-tabulated with the target variable
  - Most variables were insignificant
  - Vaccinated, dewormed, sterilized, and state variables were significant
  - The largest pets, Size 4, had a 90% adoption rate (20% over average), however, the correlation rate was low (-.023), a clue that this outlier was really due to the rarity of Size 4 pets (0.22% of the sample), not their popularity

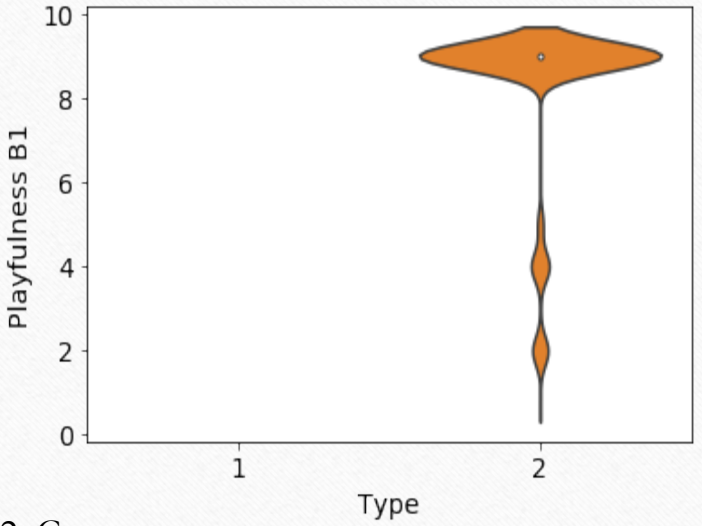
Data Wrangling

- **Outliers:** Age capped at 99% (84 months) due to implausible outliers beyond
- **Missing Value:** Some Age values were 0; we replaced these with the median age
- **Skewed Variable**: The Fee, VideoAmt, and PhotoAmt variables skewed left (towards 0), which could easily become insignificant in the model or potentially bias the estimation. A binary variable created for each skewed variable.
- **Data Quality**: Some pets had a second breed without a first breed; we ignored these pets (5 cases)
- **New Variables**:
  - Multi-level categorical variables were converted into multiple binary indicators, e.g. state, fur length, vaccination, etc.
  - Combined similar variables: use breed1 and breed2 variables to create mixed breed variable

Exploratory Data Analysis

- Dogs and Cats have different attributes and potentially attract different patterns of adopters, therefore we separated the analysis and modeling.

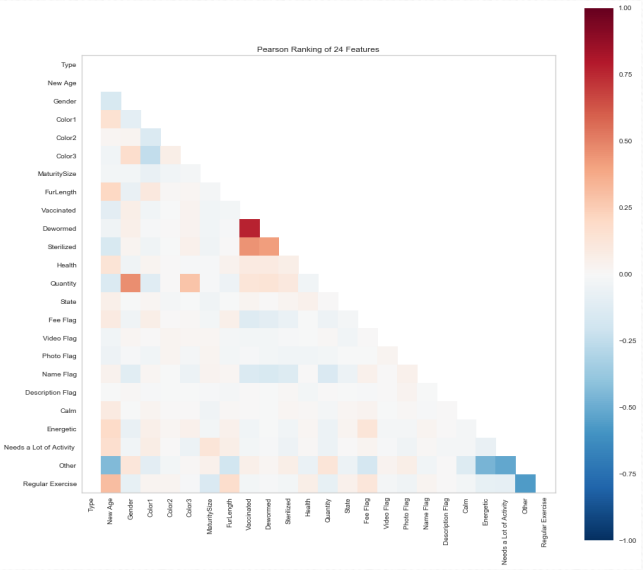


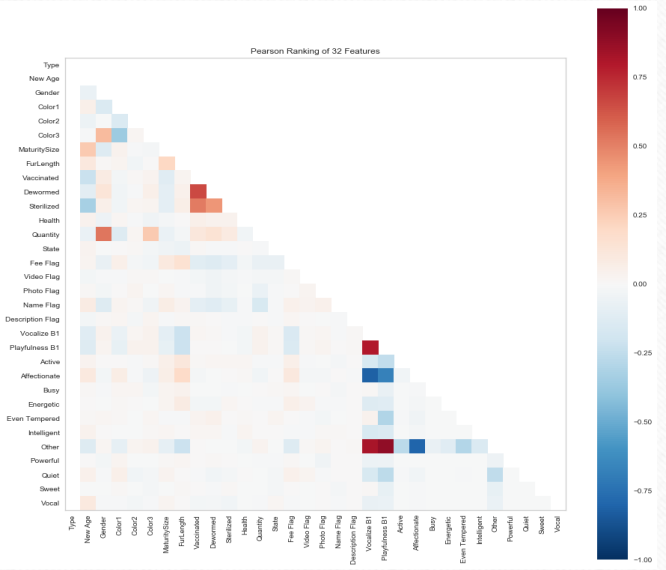


1. Dog; 2. Cat

Exploratory Data Analysis

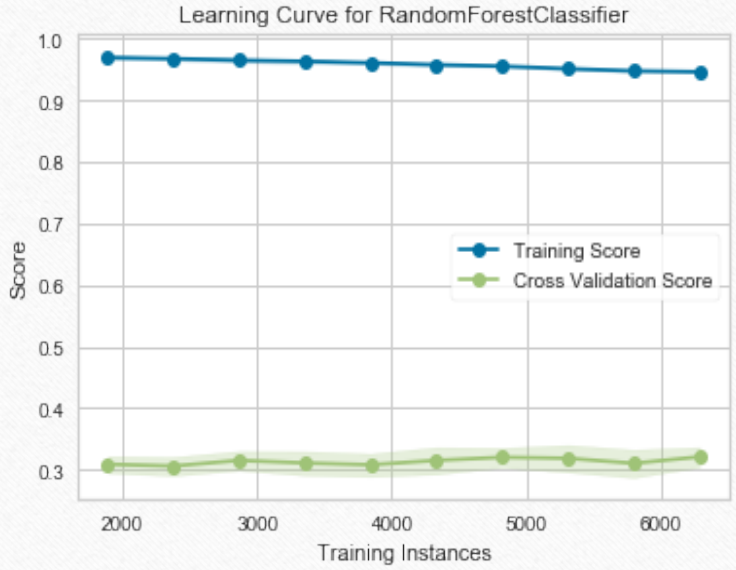
- Correlation matrix: All the explanatory variables have weak correlation with the target variable. Their correlation values range between -0.1 and 0.1

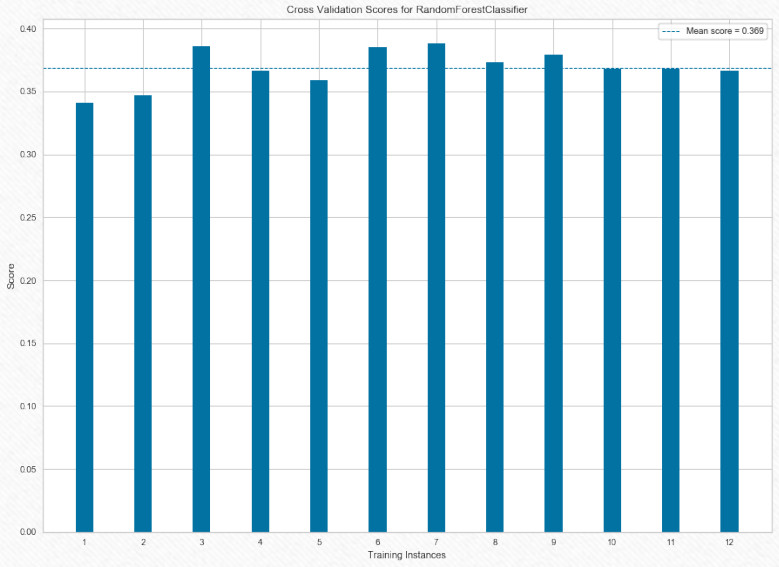




Machine Learning – Training/Testing Split

- Plot a learning curve using random forest model to explore the potential optimal training/testing data split



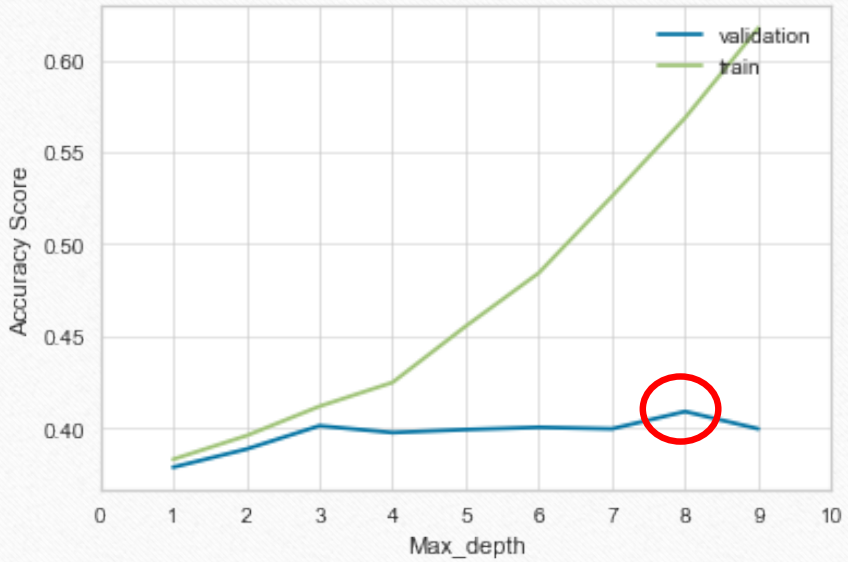


Machine Learning – Modeling Algorithm

1. Hyperparameter tuning on the primary hyperparameter in each candidate model (i.e. KNN, Random Forest, XGBoosting, SVC) to find the best hyperparameter value based on accuracy score, which is the ratio of correctly predicted observations to the total observations.
2. Calculate the classification report and feature importance plot based on the optimal hyperparameter selected from step 1 on each model. One caveat is that SVC classifier doesn't have feature importance, because the kernel function we used is RBF.
3. Run ROC plot on all the optimal models specified in Step 2.
4. Select the best classification model based on F1 score from the candidate models in Step 2.
5. Run grid search on expanded hyperparameters to further fine tune the models, and shows the detailed model parameters for our final champion models for dogs and cats, and their performance (e.g. classification report).

Machine Learning – Step 1

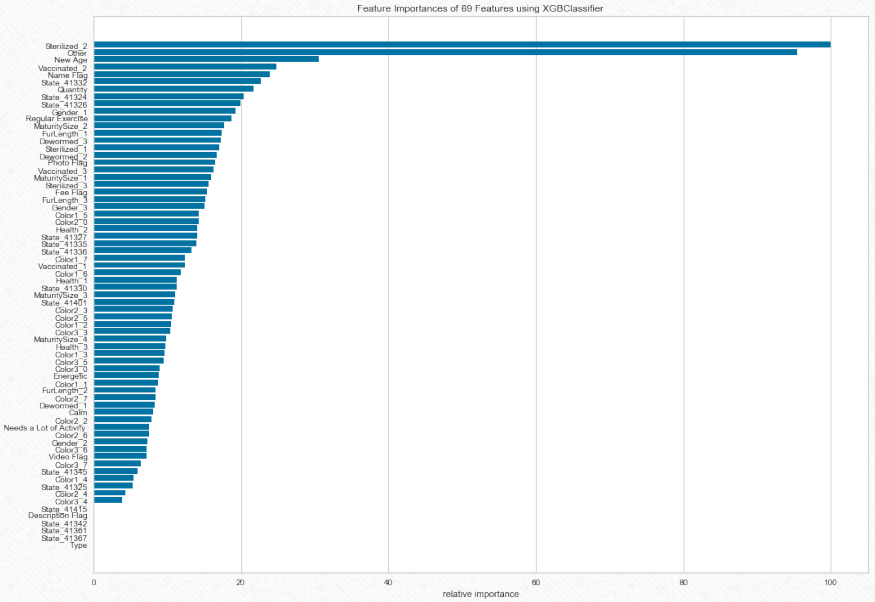
- An example of initial hyperparameter tuning on each individual candidate machine learning model – number of max depth for XGBoosting model



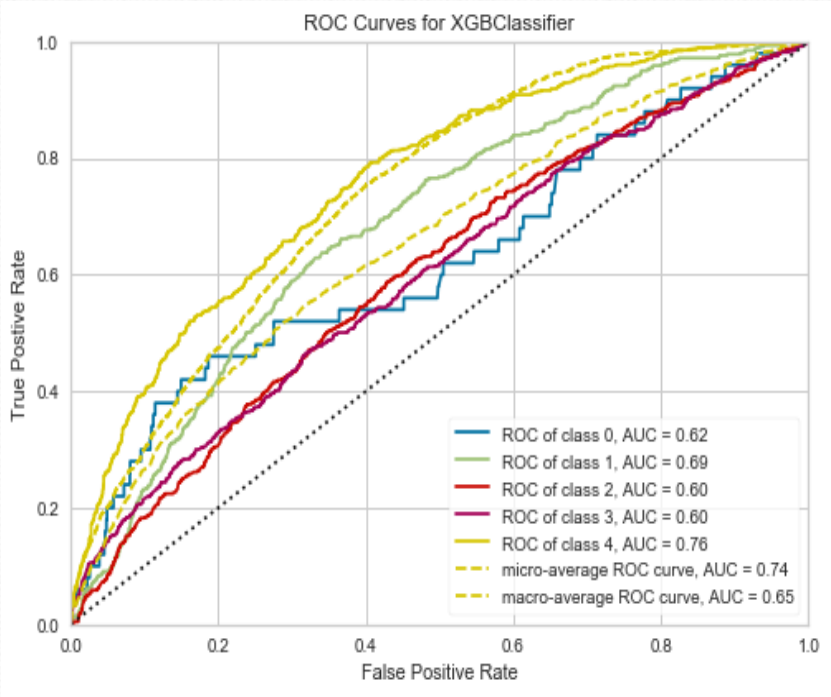
Machine Learning – Step 2

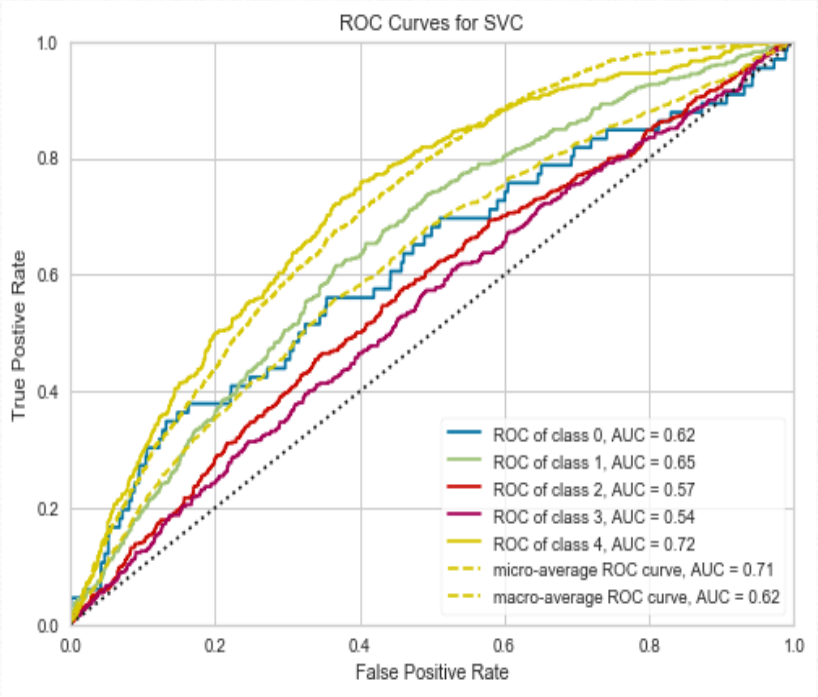
- Classification report and feature importance on optimal model after initial hyperparameter tuning. Example: XGBoosting with max depth of 8

	precision	recall	F1-score	Support
0	0.18	0.06	0.09	50
1	0.31	0.32	0.31	422
2	0.34	0.4	0.37	684
3	0.38	0.2	0.26	582
4	0.51	0.64	0.57	702
micro avg	0.4	0.4	0.4	2440
macro avg	0.34	0.32	0.32	2440
weighted avg	0.39	0.4	0.38	2440



Machine Learning – Step 3





Machine Learning - Step 4

- Select the best classification model based on F1 score (micro F1 score) from the candidate models in Step 2

Dogs	Cats
SVC: 0.3889	**SVC: 0.3706**
KNeighborsClassifier: 0.37378	KNeighborsClassifier: 0.3516
RandomForestClassifier: 0.3729	RandomForestClassifier: 0.3419
XGBClassifier: 0.3906	XGBClassifier: 0.3599

Machine Learning – Step 5

- Run randomized grid search on n_estimator, learning_rate, and subsample, max_depth, colsample_bytree, min_child_weight on XGBoosting model
- Run grid search on gamma and C on SVC

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
       colsample_bynode=1, colsample_bytree=0.9142538281542341, gamma=0,
       learning_rate=0.17308319853302173, max_delta_step=0, max_depth=6,
       min_child_weight=2, missing=None, n_estimators=377, n_jobs=1,
       nthread=None, objective='multi:softprob', random_state=0,
       reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
       silent=None, subsample=0.46955765814009887, verbosity=1)

Detailed classification report:

The model is trained on the full development set.
The scores are computed on the full evaluation set.

              precision    recall  f1-score   support

           0       0.11      0.04      0.06        50
           1       0.36      0.26      0.31       461
           2       0.33      0.37      0.35       646
           3       0.34      0.32      0.33       603
           4       0.49      0.58      0.54       680

   micro avg       0.39      0.39      0.39      2440
   macro avg       0.33      0.32      0.32      2440
weighted avg       0.38      0.39      0.38      2440
```

```
SVC(C=1, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.1, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)

Detailed classification report:

The model is trained on the full development set.
The scores are computed on the full evaluation set.

              precision    recall  f1-score   support

           0       1.00      0.03      0.06        66
           1       0.37      0.40      0.38       505
           2       0.35      0.39      0.37       585
           3       0.24      0.06      0.10       389
           4       0.41      0.59      0.48       514

   micro avg       0.37      0.37      0.37      2059
   macro avg       0.47      0.30      0.28      2059
weighted avg       0.37      0.37      0.34      2059
```

Result

Did you guess right?

Results from our model:
This pet will be adopted between 8 and 30 days (1st month) after being listed.