MATH 4995 Project 2: “Pawpularity” Prediction Using Meching Learning with Tabular Metadata and Images

Zhihao SHAO[1,2] (zshaoac@connect.ust.hk)
[1]: Department of Life Science, HKUST; [2]: Department of Mathematics, HKUST

1. Introduction

Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. Pets with attractive photos to generate more interest and be adopted faster. But what makes a good picture? With the help of data science, you may be able to accurately determine a pet photo’s “pawpularity” .

In this project, we are given both the tabular metadata and the pet images. Therefore, I would like to apply the regression methods taught in the first half of the course to tabular metadata, and computer vision-related machine learning techniques will be used to interpret the image data.

2. Dataset Description : PetFinder.my

Photo metadata:

- Manually labeling each photo for key visual quality and composition parameters

- **Focus** - Pet stands out against uncluttered background, not too close / far.
- **Eyes** - Both eyes are facing front or near-front, with at least 1 eye / pupil decently clear.
- **Face** - Decently clear face, facing front or near-front.
- **Near** - Single pet taking up significant portion of photo (roughly over 50% of photo width or height).
- **Action** - Pet in the middle of an action (e.g., jumping).
- **Accessory** - Accompanying physical or digital accessory / prop (i.e. toy, digital sticker), excluding collar and leash.
- **Group** - More than 1 pet in the photo.
- **Collage** - Digitally-retouched photo (i.e. with digital photo frame, combination of multiple photos).
- **Human** - Human in the photo.
- **Occlusion** - Specific undesirable objects blocking part of the pet (i.e. human, cage or fence). Note that not all blocking objects are considered occlusion.
- **Info** - Custom-added text or labels (i.e. pet name, description).
- **Blur** - Noticeably out of focus or noisy, especially for the pet’s eyes and face. For Blur entries, "Eyes" column is always set to 0.

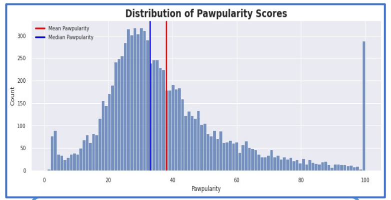
Image data:

- The training data contains 9912 pictures in total, each with a pawpularity score.

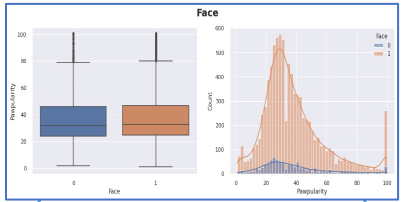
3. Exploratory Data Analysis

Photo metadata:

- The distribution of target variable – pawpularity score is slight skewed to the left, and plenty of samples have 100 score.
- The distribution of pawpularity scores is very similar for each predictor and class.



The distribution of pawpularity scores in training dataset



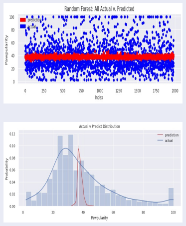
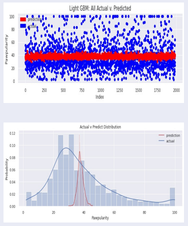
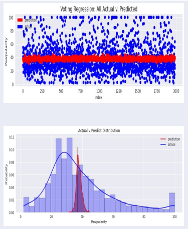
Association between ‘Face’ predictor and pawpularity score

4. Regression with Tabular Metadata

In this project, Random Forest and Light GBM were used, and then these two learners were combined with voting regression.

1) Feature engineering
    - Add normalized size and shape of images
    - Add k-means clustering results
    - Add PCA features
2) Hyperparameter tuning
    - 5 fold CV
3) Model training
4) Results

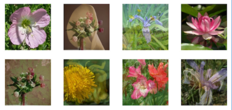
Random Forest	Light GBM
n_estimators: 50 max_depth: 5 max_features: 9	num_leaves: 3 subsample_for_bin: 30 min_child_samples: 25 min_split_gain: 0.05

Methods	Random Forest	Light GBM	Voting Regression (equally weighted)
RMSE (test split)	20.787	20.725	20.746
Prediction v.s. Actual scores			

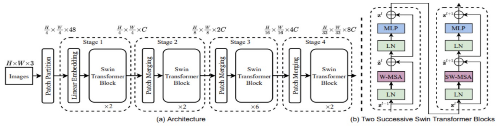
5. Swin-Transformer with Image Data

In this project, hierarchical vision transformer using shifted windows was used to leverage its unique contextual embedding advantage and the CNN-like hierarchical representation and locality.

1) Image data augmentation
    - Random flip/crop/affine, ColorJitter
    - MixU

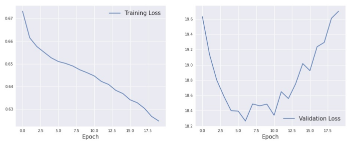


2) Model construction

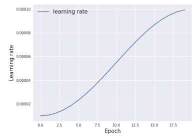


3) Model training
    - LR scheduler: CosineAnnealingWarmRestarts
    - Optimizer: AdamW
    - Train loss: BCEWithLogitsLoss
    - Max epoch: 20
4) Results: **best RMSE score in CV is 17.881**

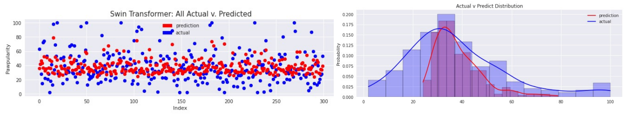
Training & Validation loss



Learning rate



Prediction v.s. Actual scores



7. Conclusion

Predictions based on image data are genuinely better than those based on binary features. Methods integrating these two data sources may maximize their prediction power.