Homework 10 (IDS21-MOUZ)

Members: Marti Orav, Uku Zingel
Repository: miOrav/IDS21-MOUZ (github.com)

Task 2 - Business Understanding

Identifying Business Goals

Background

PetFinder.my is Malaysia's leading animal welfare platform, featuring over 180,000 animals with 54,000 happily adopted. PetFinder collaborates closely with animal lovers, media, corporations, and global organizations to improve animal welfare.

Currently, PetFinder.my uses a basic Cuteness Meter to rank pet photos. It analyzes picture composition and other factors compared to the performance of thousands of pet profiles. While this basic tool is helpful, it's still in an experimental stage and the algorithm could be improved.

Predicting an accurate cuteness measure for photos helps people select the best photo when trying to appeal to potential adopters.

Business goals

This project has 2 goals.

1. Train a model to predict cuteness as accurately as possible.
2. Train models to classify optional features for pet pictures, to aid the algorithm and/or other implementations

Business success criteria

The success for the first goal will be measured through MSE.
The success for the second goal will be measured through accuracy in classifying optional features in the pictures.
Specific numerical thresholds to categorize a result as satisfactory will be decided later on, when working with the data and models.

Assessing Our Situation

Inventory of resources

The main resource for this project is a collection of roughly 10000 pet pictures, available at PetFinder.my - Pawpularity Contest | Kaggle. Additional resources include pretrained models for classifying features in pictures (e.g. for eyes).

Requirements, assumptions, constraints

- Project has to be completed by 15th December.

- Instructors must have access to the project.
- A video and project poster have to be produced
- The picture dataset cannot be shared publicly

Risks and contingencies

If the prediction of cuteness proves to be difficult through the use of classified features, we need to swap over to predicting through other means. A proper time distribution must be enforced to mitigate this risk.

Terminology

Pawpularity score - derived from each pet profile's page view statistics at the listing pages, using an algorithm that normalizes the traffic data across different pages, platforms (web & mobile) and various metrics.

Costs and benefits

The costs and benefits for this project cannot be measured through quantifying means. Instead, the cost will be mostly time (roughly 30 hours per person) and benefits will be acquired know-how and skills.

Defining Our Data-Mining Goals

Data-mining goals

Find a pre-built model for each classifiable optional feature.
Determine the best neural network type for cuteness measure regression.

Data-mining success criteria

Finding any pre-built models is a success, reducing the workload.
If the neural network type can be minimized to a choice between 2 different options, the subtask can be considered a success.

Task 3 - Data Understanding

Gathering data

Outline data requirements

Data: We require two types of data:
1. Profile pictures
2. Metadata about those pictures

Profile pictures are just that, a lot of picture files we need to analyze. The metadata about the pictures is in tabular form, and has yes/no values for a lot of different attributes.

Verify data availability

The data is available here. The page also contains a description of the dataset, including an overview of the metadata

Define selection criteria

We will be using all the pictures within the train and test folders of the project. We will also be using all the columns of the train.csv and test.csv files.

Describing Data

We have 9912 different profile pictures. We also have 9912 rows of metadata, one for each picture. The metadata includes 14 different values:

- Id - a hash value to uniquely identify pictures
- Focus - Pet stands out against an 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 a significant portion of the 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.
- Pawpularity - see terminology

The values of Id and pawpularity are the ones that do not actually describe the profile, they are there to help us handle the data and make predictions. Id binds metadata and pictures, while the pawpularity score is the evaluation of how popular a profile is, and is what we need to perform any sort of prediction in the future.

All of this data comes from Kaggle.
The data is suitable for our data-mining goals.

Exploring Data

After analyzing the metadata, I will describe what I found for each value.

Id (hash) - identificator, bind actual pictures to their metadata, not really relevant in analysis
Pawpularity (integer from 1 to 100) - this is the score, which we need all other attributes to predict.

The following attributes have a value of either 0 (false) or 1 (true):

- **Subject Focus** - This is almost always 0 in the data, which means that in most cases, this does not change pawpularity, but we need to figure out if the value being 1, improves the pawpularity score in any meaningful way. (mean 0.027)
- **Eyes** - In most cases in the data, the eyes of the pet are in the picture. (mean 0.77)
- **Face** - 90% of the profile pictures have the face of the pet. (mean 0.9)
- **Near** - Most profile pictures have the pet relatively near to the camera. (mean 0.86)
- **Action** - Very little profile pictures have the pet doing some sort of action. (mean 0.009)
- **Accessory** - Very little profile pictures have accessories in them. (mean 0.06)
- **Group** - Some profile pictures have more than 1 pet. (mean 0.13)
- **Collage** - Very little profile pictures have collage. (mean 0.05)
- **Human** - Some profile pictures also have a human in them. (mean 0.16)
- **Occlusion** - In some cases the pet is occluded. (mean 0.17)
- **Info** - A few profiles have information about them. (mean 0.06)
- **Blur** - A few profile pictures are blurry. (mean 0.07)

Verifying data quality

First I looked at the main data problems and found:

1. Data exists
2. It exists and I can have it
3. I don't find severe quality issues

After examining the data, there are no missing values, there might be some data imbalances in the metadata, but since we have 14 different attributes, this should be fine.
The pictures seem to all be relevant and on topic.

All in all we have good quality data.

Task 4 - Planning our project

First of all, we plan to use neural networks to analyze the pictures and determine a model for predicting the Pawpularity Score of new profiles. Then we plan to use the metadata with some model to help us in the prediction.

Tasks:

1. Trying out different neural networks to make a model for predicting pawpularity score.
2. Comparing the results of the previous task and selecting the best model.
3. Trying different methods for analyzing the metadata of pictures.
4. Comparing the results of the last task and selecting the best method.
5. Reflecting on the results and combining the two into the best predicting model.
6. Final touches and making the presentation.

Both of us are going to put roughly this amount of hours into each Task:

Task 1: 15 hours
Task 2: 1-2 hours
Task 3: 10 hours
Task 4: 1-2 hours
Task 5: 1-2 hours
Task 6: 2-5 hours

These numbers are our original plan, and may vary in the actual project, when we encounter unexpected difficulties.