John – Contribution summary.

Over the course of the project development cycle I have learned quite a bit of machine learning concepts and things like that. Firstly, I had no idea what I was doing. I started off with using OpenCV and Pillow which are used for image processing, but I only used them to open and load in images. I believed Tensorflow could be used to alter them and save them back out, but I didn’t properly understand how tensors worked so this first project half worked until the saving anyways.

Afterwards I tried to understand and learn the numpy and pandas libraries in Python, to process CSV data and to load in images and labels and kind of merge them and maybe eventually get a classifier going. I found a tutorial that had used folders for label names and got a script working that saved photos into individual folders based on PetID, but after a little while further into development we ended up dropping the classifier since I wasn’t sure how exactly to get QWK (Quadratic Weighted Kappa) scores that the other groups on the Kaggle leaderboard were getting. Somewhere during this process is where me and James(Seongyeop) started working more closely together by sharing ideas about how to process information and what technologies to try and use.

Lastly, I was informed by Kevin about other team’s baseline notebooks. I followed a baseline on processing sentiment data and metadata to get a feel for what it was doing with extracting features from a CSV, before trying another and incorporating some data and functions from both. We found it was effective to use Keras and a pretrained densenet (neural network) to extract numerical features from the pet images and save them into columns. This was simply a process to extract features that Keras thought was relevant with predicting and took about 8 to 12 hours to do on the entire collection of files. Near the end (around RE/ACTION day) we found that we could have included a parameter to also include more meaningful names of these features, but it was too late to predict against the entire model again. The rest of my baseline functions that I had gone through to try and understand as well was taking sentiment and metadata features as well, combining them into the image data columns and running it through the LightGBM function in which we got our QWK score and importance values. After this, I ran it through one more time, but had dropped a number of columns of features that we determined to not be as important, I worked with Kevin on this. After I saved the entire CSV with the combined data and sent it to Kevin to include in his final model so he could process it and use what he felt was needed.