Doc List

FileName	Function Description
EDA.ipynb	load up meta file and images, combine them into one dataframe EDA, remove null and unknown, save the clean dataframe to a parquet file
PCA.ipynb	reduce the feature size from 100,352 to 50, save as a parquet file
three_model_pca.ipynb	PCA on the features generated by all three neural networks.
LogisticRegression_PCA.ipynb	read pre-saved quarquet file, split data into training/ validation set by a ratio 0.8:0.2 1. fit logisticRegress with 100,352 features, 5X CrossValidation 2. fit LR mode with PCA top50, 5X CrossValidation
ResNet50.ipynb	train a new transfer learning model Base model is ResNet50, train and validate the new model, the feature size of final model is 2048
LR_resnet_only.ipynb	split data into training/ validation set by a ratio 0.8:0.2 re-train LogisticRegression model with the new transfer learning model and 5X CrossValidation
RandomForest_PCA.ipynb	split data into training/ validation set by a ratio 0.8:0.2 1. fit RandomForest model with PCA top50, 5X CrossValidation 2. fit RF model with the new TL modeland 5X CrossValidation
GBT_PCA.ipynb	split data into training/ validation set by a ratio 0.8:0.2 1. fit Gradient Boosted Tress model with PCA top50, 5X CrossValidation 2. fit RF model with the new TL modeland 5X CrossValidation

Data_Organizer.ipynb	A small script for organizing files in the image directory into relevant subdirectories. Mostly a quality of life thing, open if you want to make sure we know how to use shutils.
Image_featurizer.ipynb	Loads images into a dataframe of spark image data sources, applies models to featurize image files, and combines the files with patient metadata for analysis.
LogisticRegression_all_features	Explore 5 fold CV logistic regression on the features yielded by all three transfer learning models. As well as undersampling.
RandomForest_all_features	Explore 5 fold CV Random Forest on the features yielded by all three transfer learning models. As well as undersampling.
GBT_all_features	Explore 5 fold CV one vs rest gradient boosted trees on the features yielded by all three transfer learning models. As well as undersampling.
Benchmark_notebook_local	Train a resnet 50 transfer learning model in local mode and time it (big freaking oof). Our Rivanna work is… suboptimal.