Adrian Marinovich
Springboard Data Science Career Track
Capstone Project #1 - Milestone Report
October 12, 2018

Project: Detection of smiles in images of faces

1. Problem statement: Why it's a useful question to answer and for whom.

What is the problem you want to solve?
The problem is to detect smiles in images of faces. In other words, the problem is to build a model that will classify images of faces as either smiling or not smiling.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?
The detector developed here may find eventual implementation in a human-machine interface, as one step in developing emotional communication tools to control such things as musical instruments via MIDI, or to allow for more safe and reliable human-robot interactions, so that robots can better infer human intentions and thereby enhance their response decisions. The clients for such applications may be found in the music, robotics, and physical programming fields. A range of additional clients may be interested in smile detection as a step towards emotional classification from human facial expression, for example in market research, to better gauge interest in and reaction to products in order to guide decisions on product design, or gaming, to support decisions in creating more immersive and exciting player experiences.

2. Description of the dataset, how you obtained, cleaned, and wrangled it.

What data are you using? How will you acquire the data?
The data will be obtained from a labelled subset of the cropped version of the Labeled Faces in the Wild (LFW) dataset (LFWcrop: https://conradsanderson.id.au/lfwcrop/), in which faces are centered on the image with the background largely omitted. The LFWcrop dataset consists of 13,233 images, available as both 3-color and grayscale. The list of face images labelled as smiles comprises 600 images, and the list of face images labelled as non-smiles consists of 603 images (lists available at: https://data.mendeley.com/datasets/yz4v8tb3tp/5). This is a balanced dataset with ~600 images per target class. The cropped images have a resolution of 64x64 pixels.

The maximum dimensionality of each image is 12,288 with 3 colors (64x64x3). Limiting analysis to grayscale images would yield a reduced dimensionality, D, of approximately 4,096. This gives a D/N ratio of 6.8 (4,096/600) at the outset. Dimensionality reduction techniques, such as principal components analysis, are expected to further reduce the D/N ratio, as needed. Regularization techniques will be used to reduce overfitting. For neural networks random dropout layers will help ensure a generalized model that should work on an untrained subsample of the dataset.

What kind of cleaning steps did you perform?
The primary data wrangling tasks were not cleaning, but rather were focused on writing scripts to fill the AWS S3 buckets with data, establishing a pipeline from the Jupyter notebook to the S3 buckets using Boto, and then using the lists of smile and non-smile image file names to select the corresponding images from the larger LFWcrop dataset of images.

How did you deal with missing values, if any?
Two file names in the smile list did not have matching image files, and were identified as extraneous .txt file names ('listt.txt', and 'SMILE_list.txt') which did not interfere with the correct matching of image file names.

Were there outliers, and how did you handle them?
Images corresponding to the smile and non-smile lists were reviewed to check for validity of the smile/non-smile labels, and found to be consistent with the given labels. There also appears to be a diversity of age, sex, ethnicity, head position, facial hair, and presence of eyeglasses in both the smile and non-smile sets, with the exception that children appear to be absent from both sets. Given that children's photos would not be used without careful consideration and protections, this is not seen as a flaw or outlier issue in the dataset, for the purposes of an initial analysis as undertaken here.

Additional wrangling:
The data were split into train and test datasets, with 1000 images in the train dataset, and 203 images in the test dataset.

To remove possible index-specific biases embedded in the data, a randomly shuffled index was created and applied to both the feature (pixels) and target (labels) datasets.

3. Initial findings from exploratory analysis

3a. Summary of findings

The exploratory data analysis (EDA) was conducted after an initial machine learning approach, random forest, was performed to classify the images into smile and non-smile categories. Before the random forest was performed, a preliminary comparison of Gini vs. entropy impurity criteria was performed using a lone decision tree classifier, with entropy criterion then selected to be used in the random forest classifier. The EDA involved colormapping pixel feature importances and reviewing at a contingency table in conjunction with that.

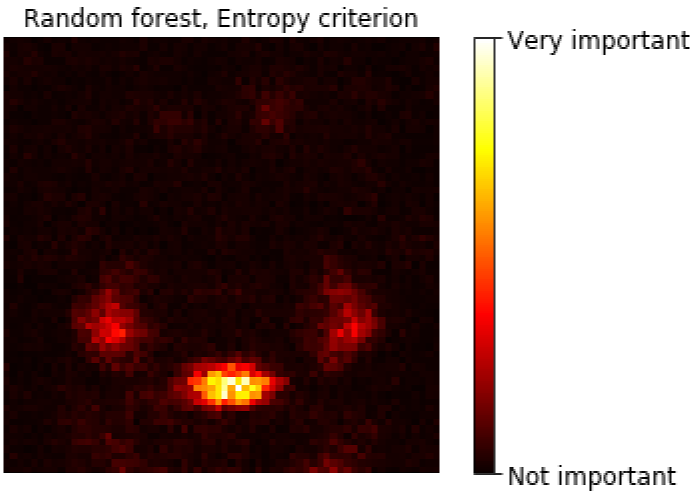
As seen in the images below, as well as the EDA Jupyter notebook referenced below, the feature importances from the random forest classifier colormapped to a plot of pixel location show a clustering of high pixel importances in the mouth and cheek regions, as well as fainter clusters in the forehead and inferolateral cheek regions. Very faint signal is also seen in inferior nose/nostril and nasolabial folds regions. This mapping of pixel importances, coupled with plotting the results of classifying the test faces on a contingency table, suggests some higher-level features that may have a role in random forest misclassification:

- False negatives may in part result from face rotation, scaling and centering/cropping differences.
- False positives may in part result from prominent nasolabial folds (as might be seen in smirking or grimacing), presence of facial hair, as well as face rotation, scaling and centering/cropping differences.

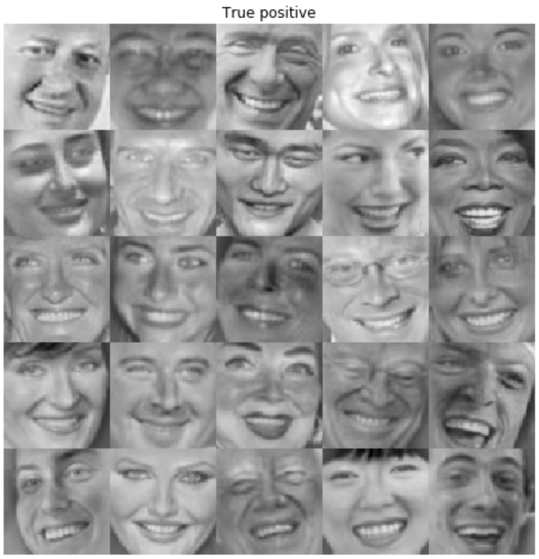
3b. Visuals and statistics to support findings

A face image example (labelled as smiling):

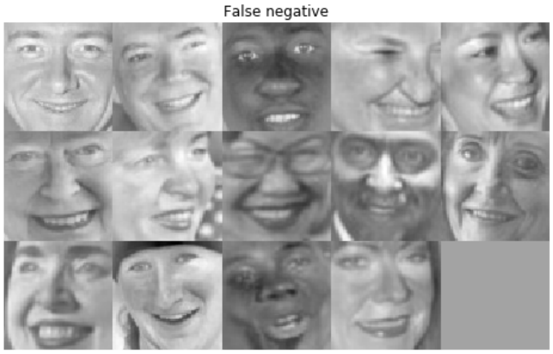
The mapping of pixel (feature) importance using the random forest smile classifier, after training on 1000 face images:



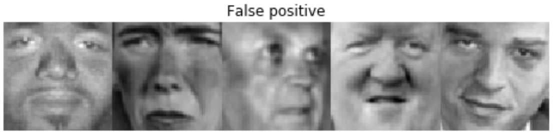
A contingency table showing images from the test set comparing random forest smile classification against the target labels.



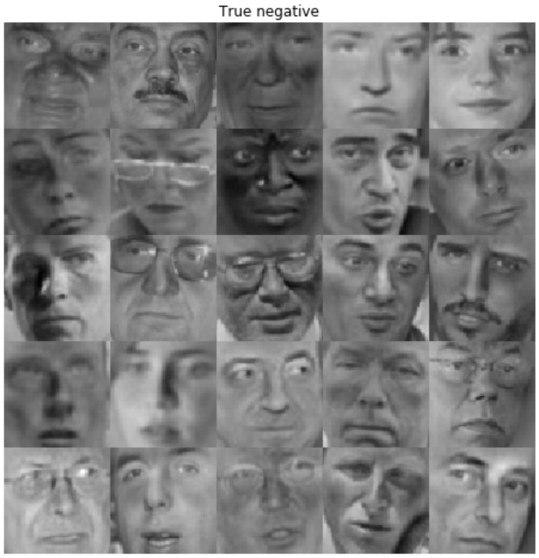
True positive



False negative



False positive



True negative

Future directions:

In the next machine learning steps involving support vector machines and neural networks, the important pixel regions may change, and with that the types of misclassification seen. This overall approach, however, will allow ongoing monitoring and adjustment of the models, as well as guide possible transformations and filtering of the image data to improve smile detection.

Links:

https://github.com/adriatic13/springboard/blob/master/dsct_capstone1/Adrian_Marinovich_Cap1_smiles_data_wrangling.ipynb

https://github.com/adriatic13/springboard/blob/master/dsct_capstone1/Adrian_Marinovich_Cap1_smiles_eda.ipynb

https://github.com/adriatic13/springboard/blob/master/dsct_capstone1/Marinovich_Cap1_Milestone_slides.pdf

References:

Arigbabu, Olasimbo Ayodeji, et al. "Smile detection using hybrid face representation." Journal of Ambient Intelligence and Humanized Computing (2016): 1-12.

Huang GB, Mattar M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report.