Cal Flett and Matt Timoney - Mattson Nutrition Reference Document

(i) null values in 'dob' and 'gender' fields were dropped entirely; this decision was made because
(ii) the case background describes these nulls as 'significant omissions'. This reduces our number of total
(iii) observations to 66,161 (-17055 or 20.49%)
(iv) 'Big 5' Categories were dummy-coded, resulting in 25 new fields broken down by response (1-5)
(v) boolean expression is then converted to a binary, then the field is added to the features list
(vi) The BMI field was uniquely challenging because of the floats and nulls. We created weight bins to solve this
(vii) Our Principle Component Analysis with five components allows us to reduce our features significantly
(viii) PCA accuracies: '12.0%', '8.1%', '5.2%', '3.8%', '2.9%'