I. TESTS FOR FEATURES AND DATA
Data 1: Feature expectations are captured in a schema
Data 2: All features are beneficial
Data 3: No feature's cost is too much
Data 4: Features adhere to meta-level requirements
Data 5: The data pipeline has appropriate privacy controls
Data 6: New features can be added quickly
Data 7: All input feature code is tested

II. TESTS FOR MODEL DEVELOPMENT
Model 1: Every model specification undergoes a code review and is checked in to a repository
Model 2: Offline proxy metrics correlate with actual online impact metrics
Model 3: All hyperparameters have been tuned
Model 4: The impact of model staleness is known
Model 5: A simpler model is not better
Model 6: Model quality is sufficient on all important data slices
Model 7: The model has been tested for considerations of inclusion (ethical concerns)

III. TESTS FOR ML INFRASTRUCTURE
Infra 1: Training is reproducible
Infra 2: Model specification code is unit tested
Infra 3: The full ML pipeline is integration tested
Infra 4: Model quality is validated before attempting to serve it
Infra 5: The model allows debugging by observing the step-by-step computation of training or inference on a single example
Infra 6: Models are tested via a canary process before they enter production serving environments
Infra 7: Models can be quickly and safely rolled back to a previous serving version

IV: MONITORING TESTS FOR ML
Monitor 1: Dependency changes result in notification
Monitor 2: Data invariants hold in training and serving inputs
Monitor 3: Training and serving features compute the same values
Monitor 4: Models are not too stale
Monitor 5: The model is numerically stable
Monitor 6: The model has not experienced a dramatic or slow-leak regressions in training speed, serving latency, throughput, or RAM usage
Monitor 7: The model has not experienced a regression in prediction quality on served data

SCORING:

- For each test, half a point is awarded for executing the test manually, with the results documented and distributed
- A full point is awarded if there is a system in place to run that test automatically on a repeated basis
- Sum the score for each of the 4 sections individually
- The final ML Test Score is computed by taking the minimum of the scores aggregated for each of the 4 sections. We choose the minimum because we believe all four sections are important, and so a system must consider all in order to raise the score.