PROJECT 3

PARKINSON'S DISEASE VOICE DATA

ERIC HANSEN

FLATIRON DS; SELF PACED

3/16/2022

BACKGROUND - PROBLEM

- Parkinson’s Disease affects seven million people globally; voice features are used for diagnosis, but can they be used more effectively?

BACKGROUND - STAKEHOLDERS

- Medical professionals, Patients With Parkinson’s (PWP), aging adults

BACKGROUND – DATA

- 'Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings', IEEE Journal of Biomedical and Health Informatics, vol. 17(4), pp. 828-834, 2013. https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings#

MODEL

- Eight non-linear classification tools, including logistic regression, k-nearest neighbors, random forest, bagging, gradient boosting and more

GOALS

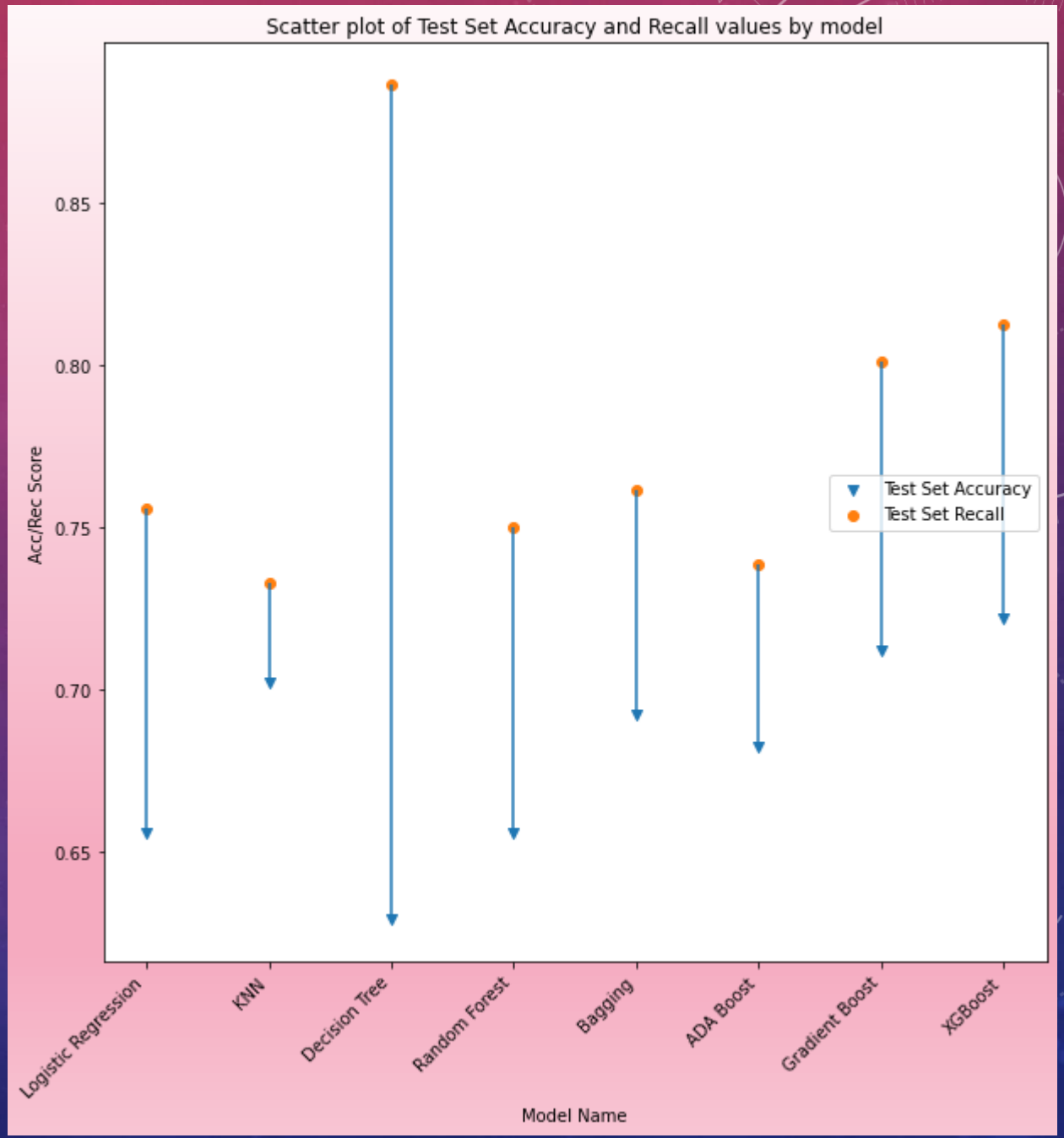
- Goals
  - Identify actionable features of voice samples that may predict Parkinson’s to augment existing diagnosis tools
  - Create optimized model for classification of new voice data

METHODOLOGY DETAILS

- 1) I performed Exploratory Data Analysis to identify characteristics of the data
- 2) I created several new features based on investigation of PD and discussion with geriatric speech pathologist, including monotonicity, oral festination, projected gender
- 3) I created numerous models and optimized their parameters for best accuracy (generally desirable) and recall score (due to priority of avoiding false negatives for PD) across cross validation
- 4) I tested each of these models on the test set
- 5) I identified a model which scored highly on the training and test set and extracted meaningful important features
- 6) I created a pipeline and pickled it for portability and ease of future use
- 7) The conclusions: XGBoost model had high accuracy and recall; it assigned high feature importance to some jitter and pitch features
- 8) Furthermore, the model is ready to accept any new voice data, to aid in diagnosis

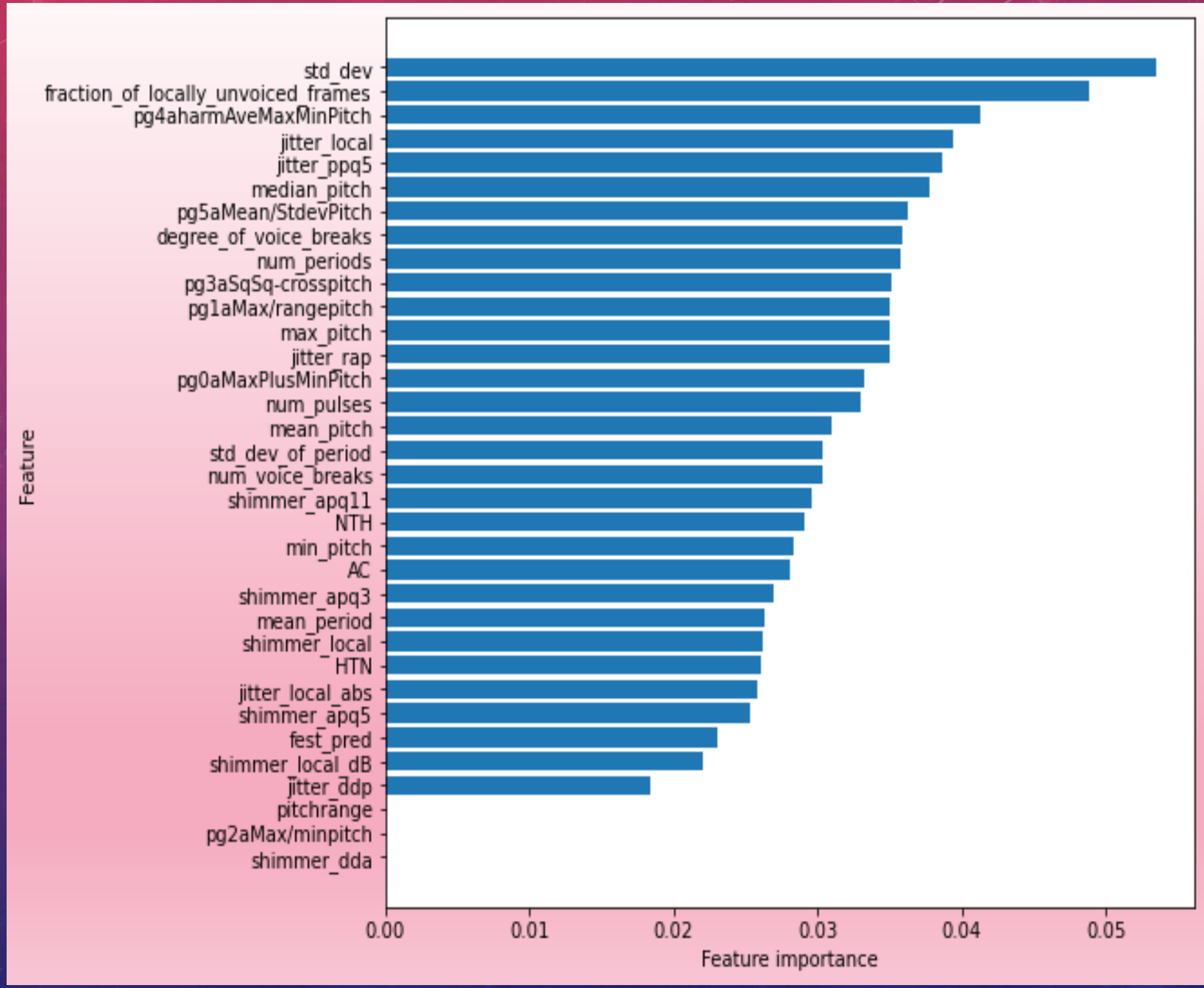
RESULTS 1 – TEST SET ACCURACY&RECALL BY MODEL

- Multiple crossvalidation rounds were taken for each model, average accuracy and recall were computed across these rounds; then tuned models tested on the test set
- One model which did well on both was XGBoost, which had 72% accuracy and 81% recall
- Other models and scores(acc, rec) are:
  - logistic regression: 66%, 76%
  - knn: 70%, 73%
  - decision tree: 63%, 89%
  - random forest: 66%, 75%
  - bagging: 69%, 76%
  - adaboost: 68%, 74%
  - gradientboost: 71%, 80%



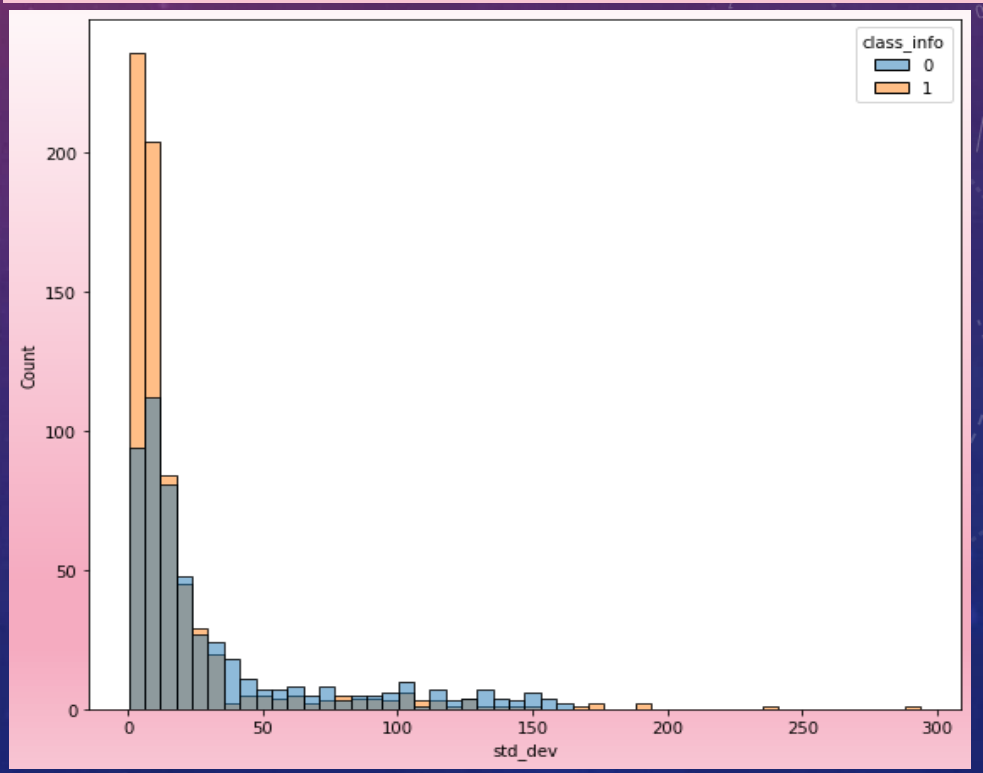
RESULTS 2 – FEATURE IMPORTANCES FOR MODEL

- The top 10 features of highest importance are:
  - Standard deviation of pitch
  - Fraction of locally unvoiced frames
  - Pg4 (harmonic avg of max/min pitch)*
  - Jitter_local
  - Jitter_ppq5
  - Median_pitch
  - Pg5(mean/stddev of pitch)*
  - Degree of voice breaks
  - Number of periods
  - Pg3 (maxpitch^2+minpitch^2-2maxp*minp)*
- * These features were manually engineered



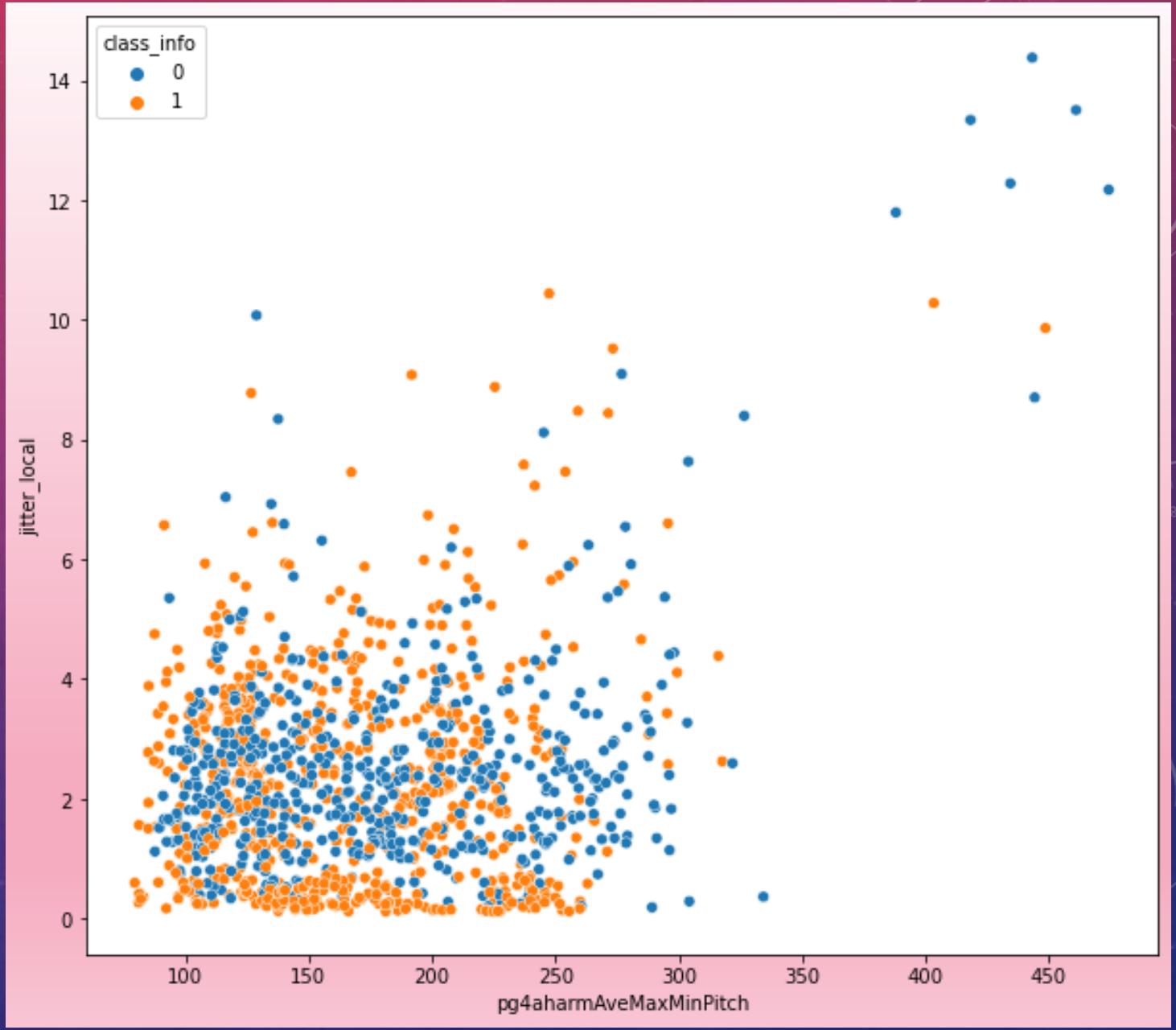
STD_DEV FURTHER INSPECTION

- Class 0 = Non PD, Class 1 = Patient with PD
- It is challenging to visually identify relationships between a numeric predictor and a binary target variable, as seen in the top graph.
- But we can observe some separation in behavior of the two classes in the bottom histogram – patients without PD appear to have more standard deviation of pitch – i.e. this may be consistent with less monotonicity



TWO FEATURE SEPARATION

- If we try to separate behavior by not just one predictor, but two, and color code by target variable, we can get a little more visual insight.
- Some combinations of important features show more separation, some show less – one example of illustrative graph is shown
- In this case, we can see that Patients with Parkinsons (class1) frequently have very low, and occasionally higher than usual jitter_local



CONCLUSIONS/QUESTIONS

- A XGBoost model has identified some characteristics to dial in on.
  - As stated earlier, the top 5 features are Standard deviation of pitch, Fraction of locally unvoiced frames, Pg4 (harmonic avg of max/min pitch), Jitter local, Jitter ppq5
  - This model had a test set accuracy of 72% and recall of 81%

CONCLUSIONS/QUESTIONS

- Actionable items:
  - Prioritize identifying speech pathologies that correspond to pitch standard deviation, fraction of unvoiced frames, and jitter.
  - Use XGBoost or other model to add layer of additional screening to PD evaluations

CONCLUSIONS/QUESTIONS

- Potential further steps:
  - acquire more labeled data to further train models.
  - Perform study over time to identify early-changing features that predict eventual PD.

CONCLUSIONS/QUESTIONS

- Questions?