PROJECT 3

PARKINSON’S DISEASE VOICE DATA

ERIC HANSEN

FLATIRON DS; SELF PACED

12/23/2021

BACKGROUND

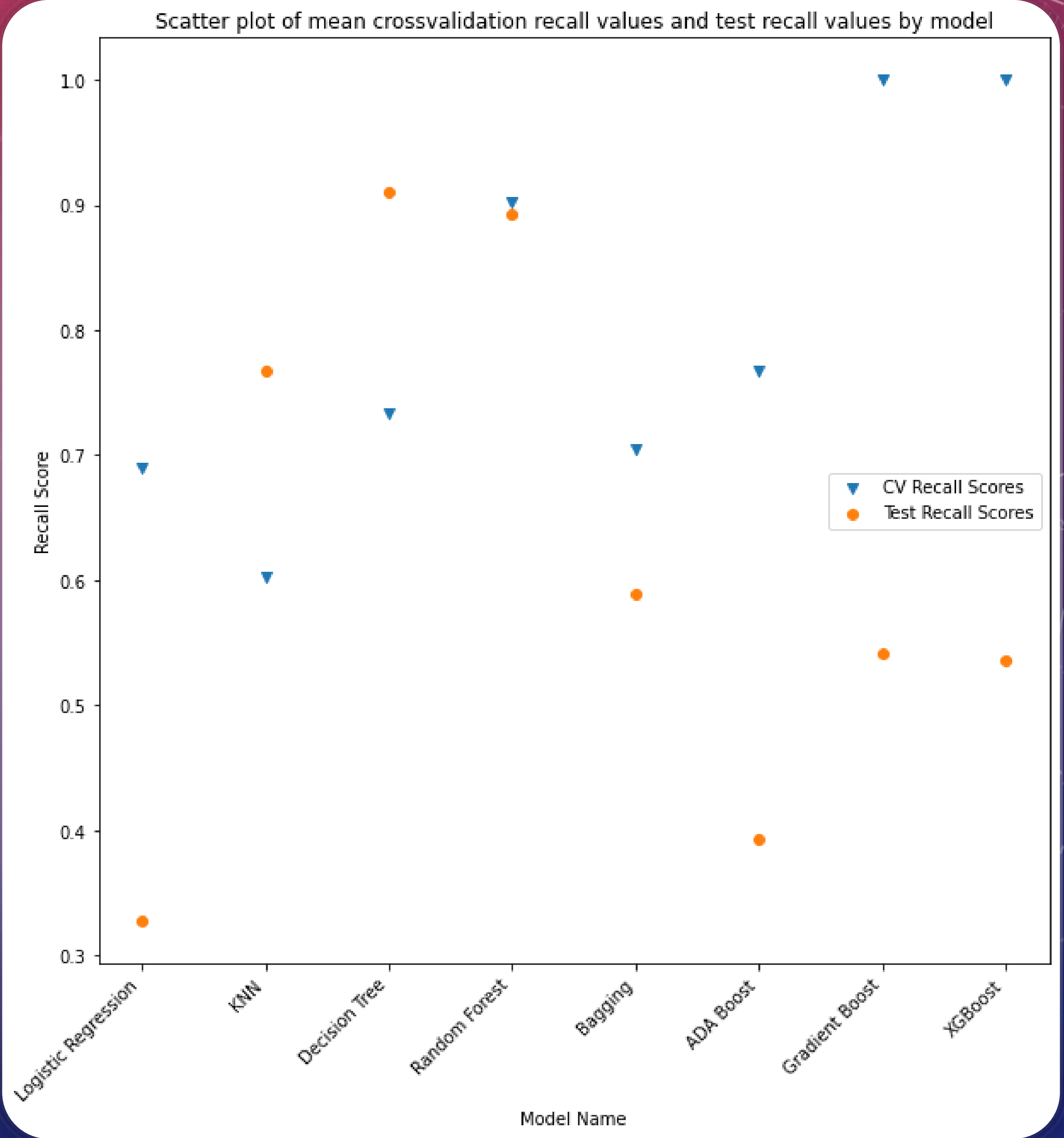
- Problem – Parkinson’s Disease affects seven million people globally; voice features are used for diagnosis, but can they be used more effectively?
- Stakeholder – Medical professionals, Patients With Parkinson’s (PWP), aging adults
- Data – 'Collection and Analysis of a Parkinson Speech Dataset with 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
  - 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

- 1) I performed Exploratory Data Analysis to identify characteristics of the data
- 2) I created several new features including monotonicity, oral festination, projected gender
- 3) I created numerous models and optimized their parameters for best recall score (due to priority of avoiding false negatives for PD) across crossvalidation
- 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: x model had a high recall score; it assigned high feature importance to:
- Furthermore, the model is ready to accept any new voice data, to aid in diagnosis

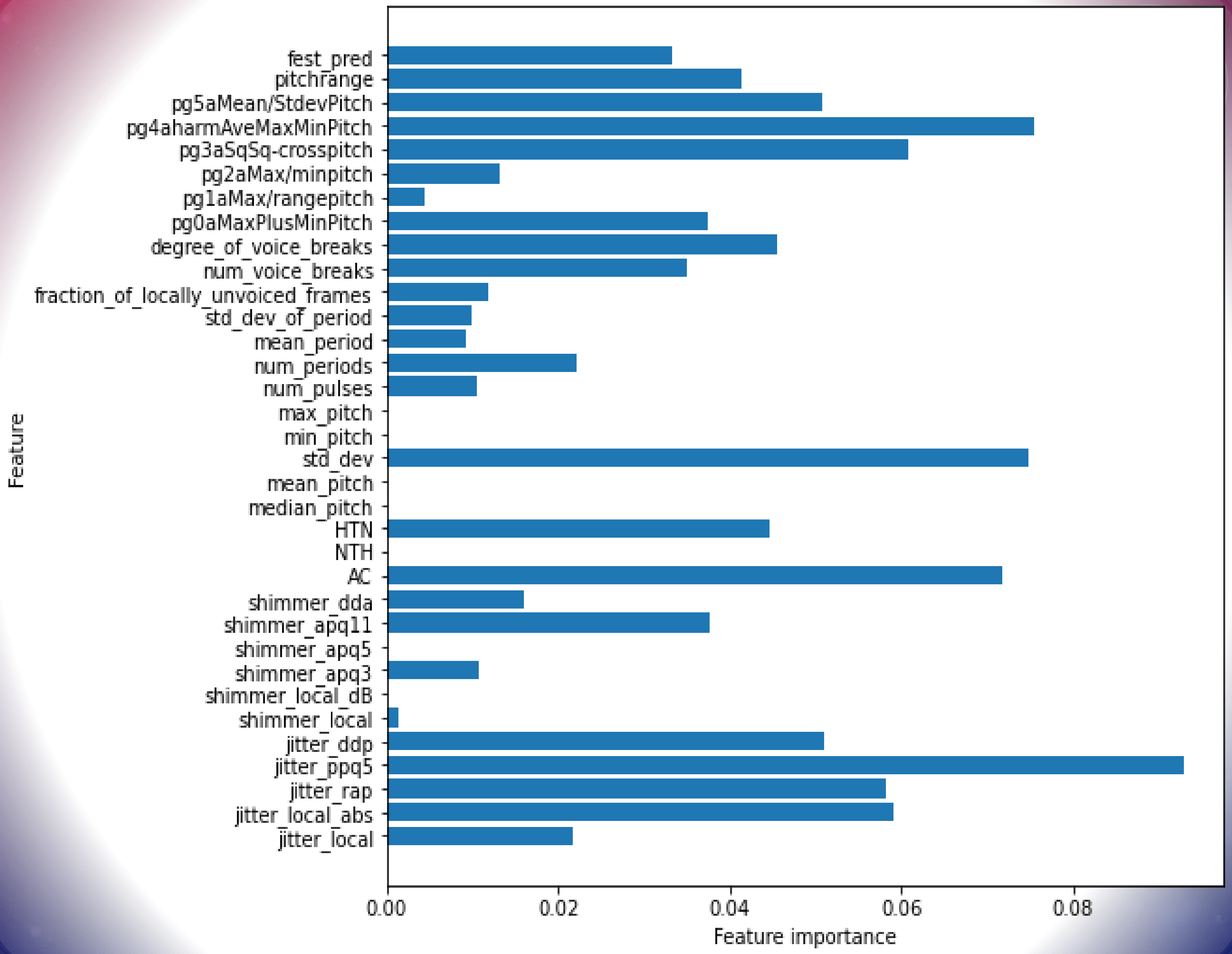
RESULTS 1 – CROSSVALIDATION RECALL VS MODEL

- Multiple crossvalidation rounds were taken for each model, average recall was computed across these rounds
- Recall was also computed for each model on the test set (which is 100% patients with Parkinson's)
- Some models had significantly different trainCV/test recall
- One model which did well on both was Random Forest, which had recall scores of 90% and 89%
- Other models and scores are:
  - logistic regression: 69%, 33%
  - knn: 60%, 77%
  - decision tree: 73%, 91%
  - bagging: 70%, 59%
  - adaboost: 77%, 39%
  - gradientboost: 100%, 54%
  - xgboost: 100%, 54%



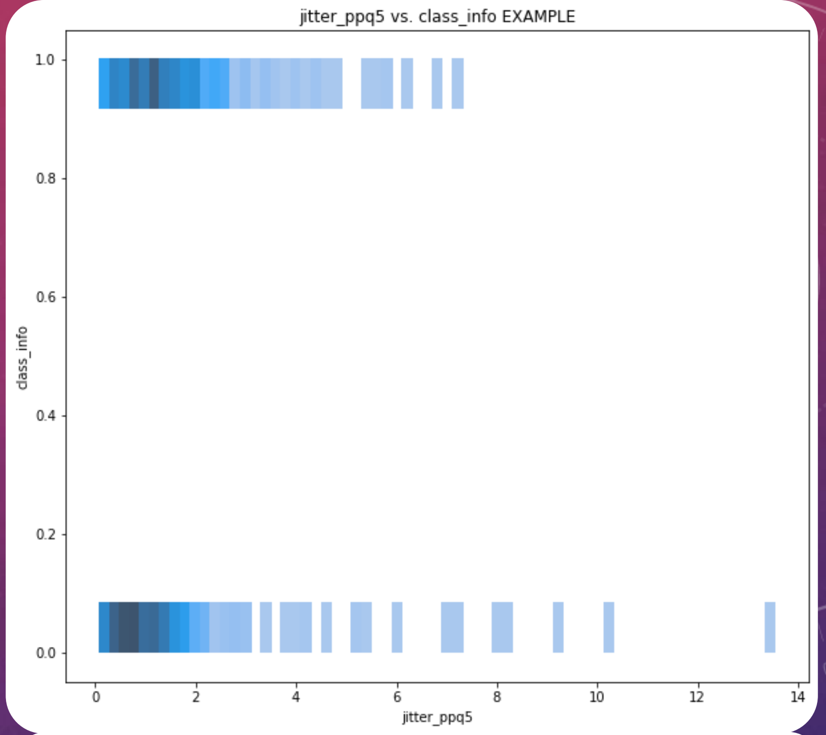
RESULTS 2 – FEATURE IMPORTANCES FOR RANDOM FOREST MODEL

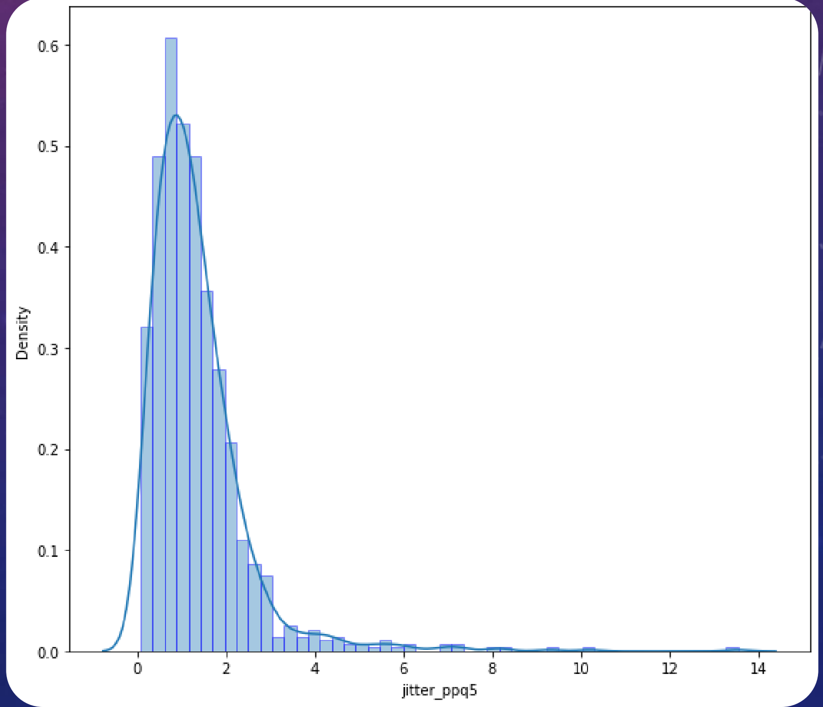
- The top 10 features of highest importance are:
  - Jitter_ppq5
  - pg4aharmAveMaxMinPitch
  - Std_dev (of pitch)
  - AC (autocorrelation)
  - pg3aSqSq-crosspitch
  - Jitter_local_abs
  - Jitter_rap
  - Jitter_ddp
  - pg5aMean/StddevPitch
  - Degree of voice breaks
- Festination and Pitch Range were in the next six



JITTER_PPQ5 FURTHER INSPECTION

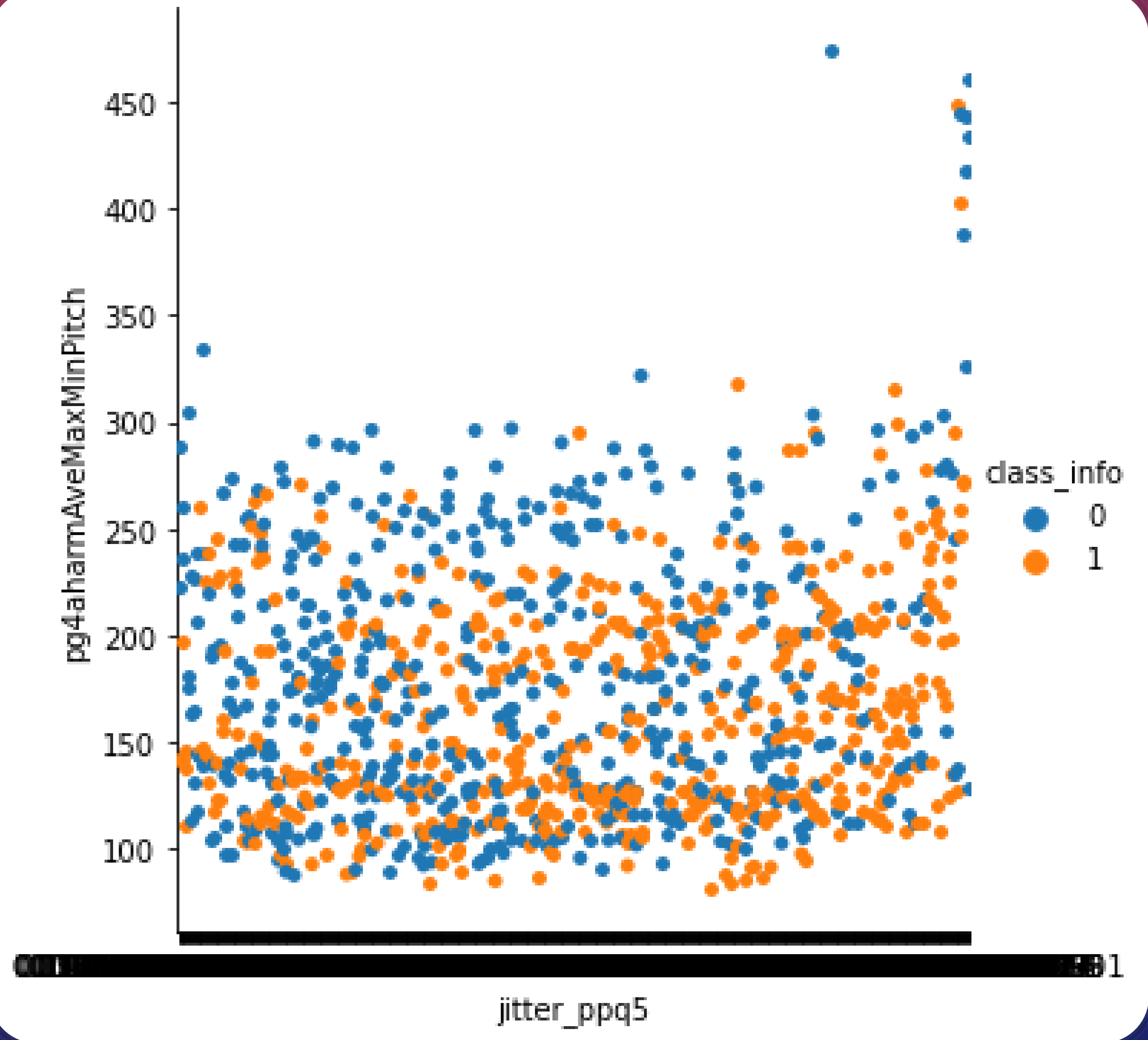
- It is challenging to visually identify relationships between a numeric predictor and a binary target variable, as seen in the top graph.
- Jitter_ppq5 didn’t have any distinctive features in its distribution function





JITTER_PPQ5 – INCLUDING PG4A

- If we try to separate behavior by not just one predictor, but two, and color code by target variable, we can get a little visual insight.



CONCLUSIONS/QUESTIONS

- A random forest model has identified some characteristics to dial in on.
  - As stated earlier, the top 5 features are Jitter_ppq5, pg4aharmAveMaxMinPitch, Std_dev (of pitch), AC (autocorrelation), pg3aSqSq-crosspitch
  - This model had a mean recall score over cross-validation of 90%. It had recall score on the test set of 89%.
- Actionable items:
  - Prioritize identifying speech pathologies that correspond to jitter_ppq5, harmonic average of max and min pitch, pitch standard deviation, and autocorrelation
  - Use random forests or other model to add layer of additional screening to PD evaluations.
- Potential further steps:
  - acquire more labeled data to further train models.
  - Perform study over time to identify early-changing features that predict eventual PD.
- Questions?