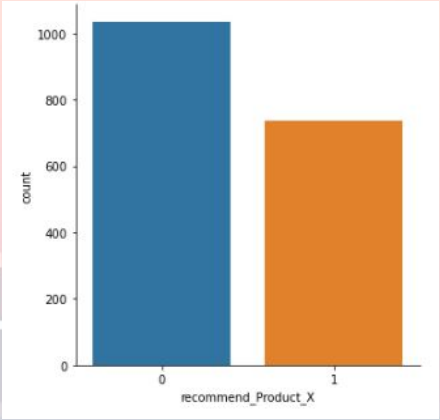
DisY Boost

Predicting physician recommendations of Product X for the treatment of Disease Y

Alex McDaniel

Problem + Data Overview

Problem:
Predict whether physicians for a given patient will recommend “Product X” to treat Disease Y



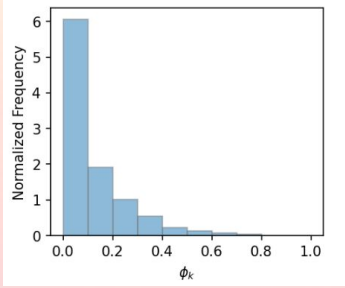
Summary of the “Chart Review Data”

- 1773 patient records, 84 features, mostly binary
- General Patient info (ID, various ages, gender, etc.)
- Disease Y severity at diagnosis vs current
- Risk factors: r1-r6 (>90% of r6 is missing, r5 not binary)
- Current diagnosis status: dx1-dx21, and history for dx22/dx33
- Current/previous treatment: px1-px12 (~60% available for previous)
- Target variable ‘recommend_Product_X’, 0 or 1
    - Slightly unbalanced dataset, ~60% no recommendation ~40% recommended

Exploring the Data

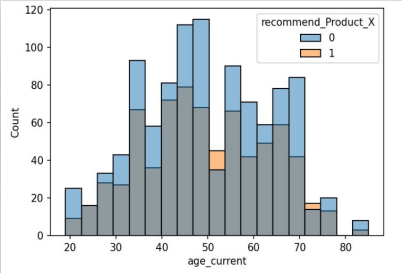
Did not observe “insightful” correlations between parameters (from correlation matrix)

Using ϕ_k metric, Baak+2019. Similar to pearson coefficient, closer to 1 more correlated.



No obvious correlation of age features to recommendations

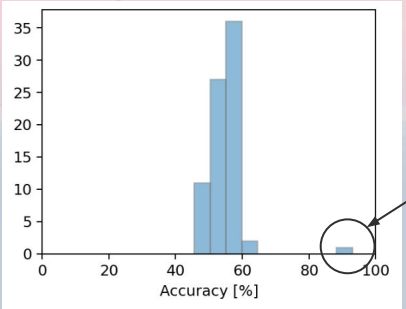
Ex. Current age



Risk Factor 6 (rf_r6_present) appears to be an excellent predictor!

Using RF6 alone gives ~93% accuracy.
(But the sample size is small (n=127))

Accuracy for single feature predictions
E.g. has_had_surgery = reccommend_Product_X



rf_r6_present

rf_r6_present provides ~93% accuracy alone.
All others ~45-60% accuracy alone.

DisY Boost

DisY Boost is built using XGBoost library.

XGBoost (e**X**treme **G**radient **B**oosting) Boosted Decision Tree Ensemble Classifier
Why use it?

- Fast, easy, powerful
- Handles null values (think 'rf_r6_present')

The Baseline model is very simple: we only label encode the categorical data and drop patientID

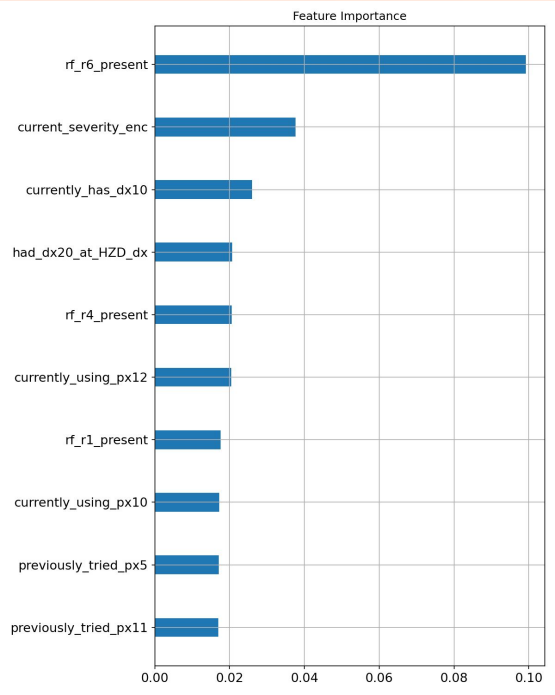
Baseline model accuracy ~86.7 ± 2.2%

Tested Adjustments to Baseline:

- Analyze feature importance
  - No improvement by dropping low importance features
- Hyperparameter Tuning
  - Indications that could improve accuracy by +Δ1-2% (But would need further testing)
- Some minor feature engineering did not improve the model
  - Very minimal, 6 new columns: sums of risk factors, dx at HZD, current dx;
  - clusters of the same

Feature Importance

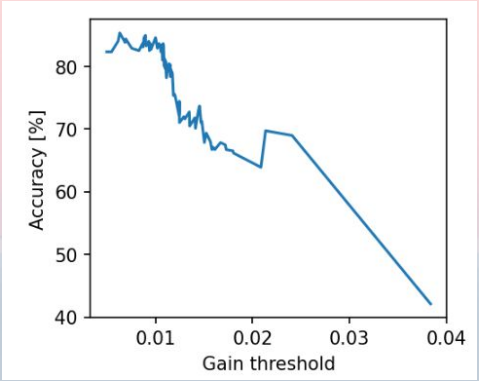
Top ten features by gain importance



Gain

Gain: improvement in accuracy brought by a feature to the branches it is on.

*Note: No model improvements by dropping low importance features.



Accuracy as features below gain threshold are removed

Final Results

We have developed a the DisY Boost model using XGBoost in order to predict whether physicians are likely to recommend “Product X” for treatment of Disease Y for a given patient.

The final DisY Boost model achieves an accuracy of 86.5 ± 2.2%

Risk factor 6 (rf_r6_present) is potentially a highly effective predictor of whether Product X will be recommended, however less than 10% of patients have this information.

The dataset and model could be significantly improved by obtaining more rf_r6_present info