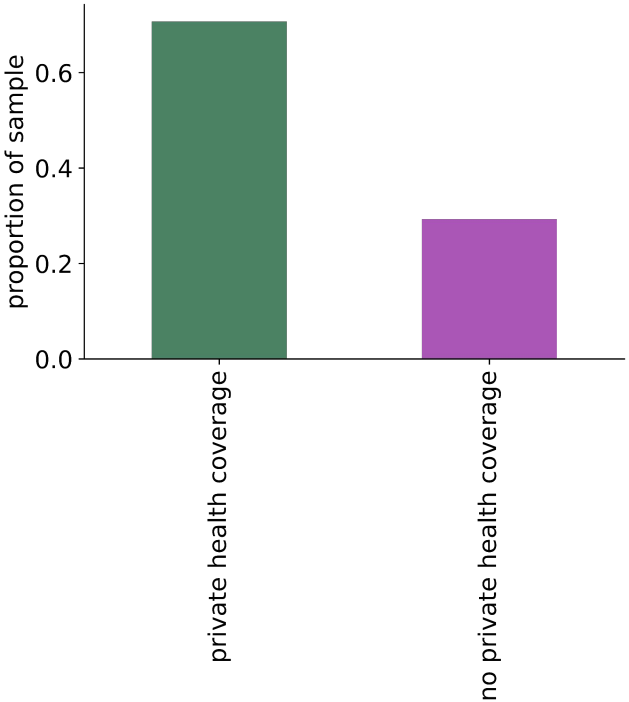
predicting private health coverage in the US

backdrop: health coverage in the US

health insurance: public or private options

most individuals have private health insurance

- through employer or labor union
- direct from company
- through government marketplace



backdrop: health coverage in the US

opportunity: private health insurance companies may increase subscribers to their plans by targeting individuals already in the market space

which factors are common to those with private health coverage?

can we predict whether an individual has private health coverage based on demographic data?

demographic dataset

US Census Bureau conducts the American Community Survey annually

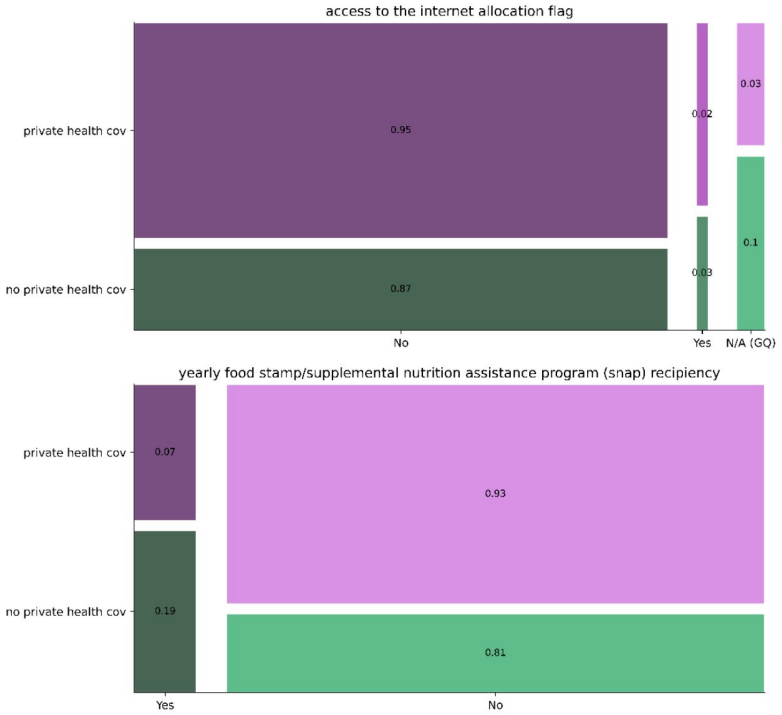
publicly available for download or through their API

info included: health coverage, income, household, employment, +

exploring the data

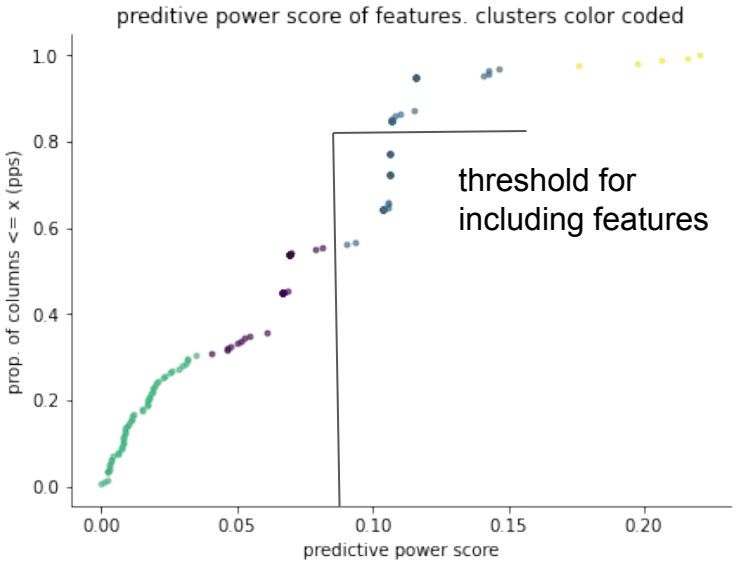
vast majority with private coverage have internet (fewer without private coverage have internet)

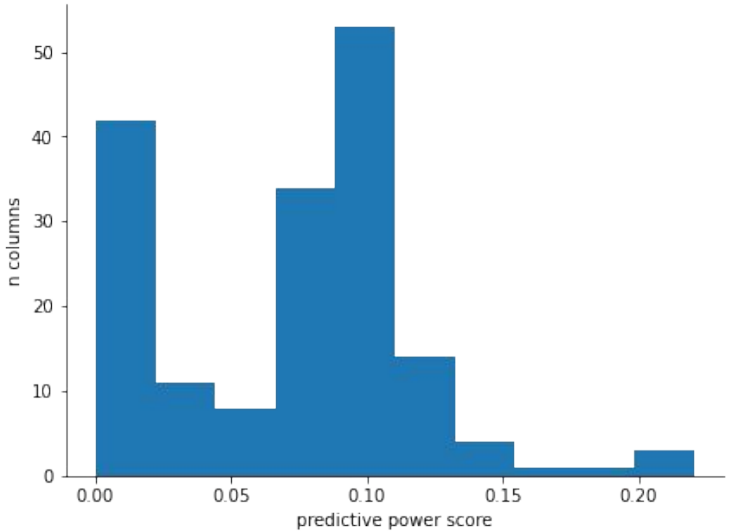
very few individuals with private coverage receive food stamps (compared to those without private coverage)



feature selection

predictive power scores and feature selection: 71 categorical, 6 numeric variables





model development

five classification models trained and evaluated:

- logistic regression
- K-nearest neighbors (KNN)
- support vector machine (SVM)
- random forest
- extreme gradient boosting (XGBoost)

model development

five classification models trained and evaluated:

- logistic regression
- K-nearest neighbors (KNN)
- support vector machine (SVM)
- random forest
- extreme gradient boosting (XGBoost)

PCA on numeric variables to reduce dimensions

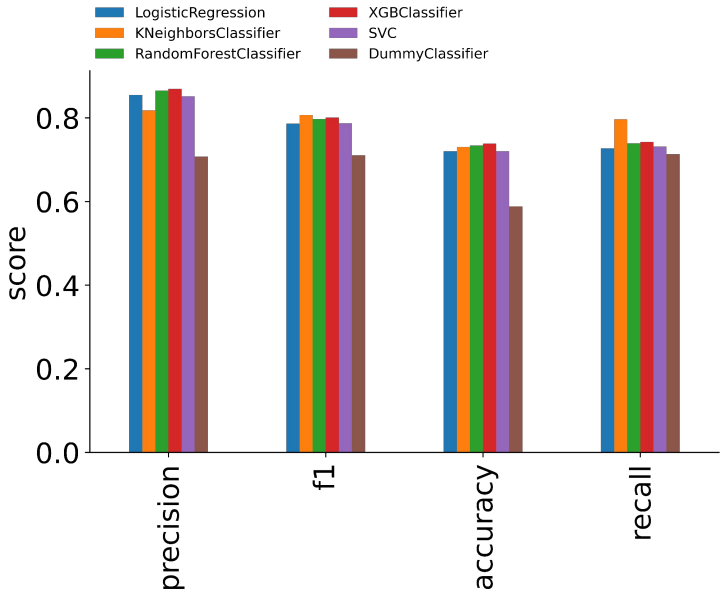
random under-sampling to counteract class imbalance

model metric comparisons

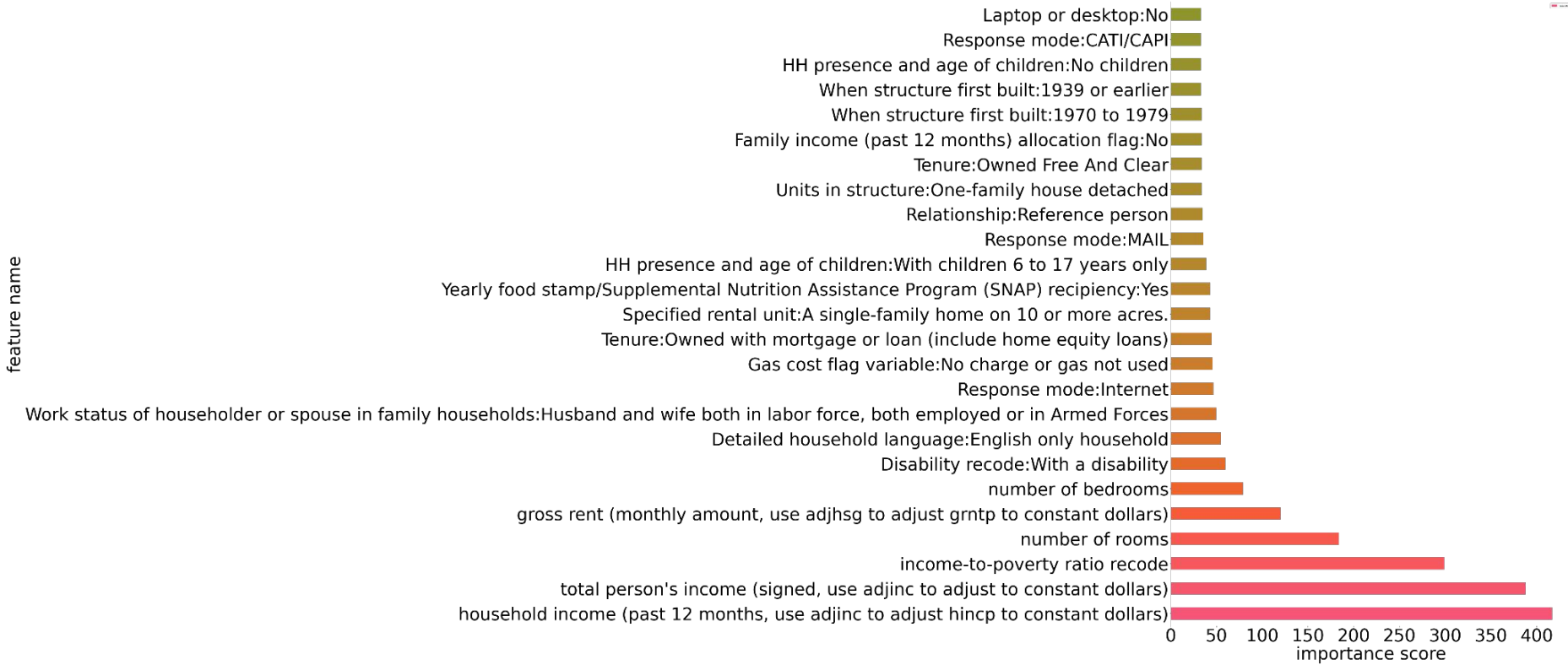
prioritize precision score: limits marketing to those already in private health coverage space

XGBoost model had highest precision score

and highest accuracy!



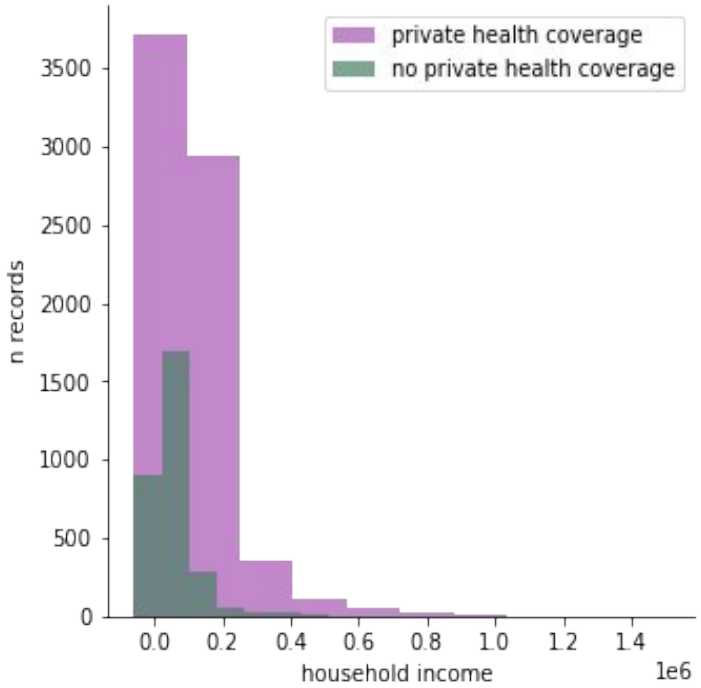
XGBoost important model features



recommendations based on model features

identify households with an income > $100,000

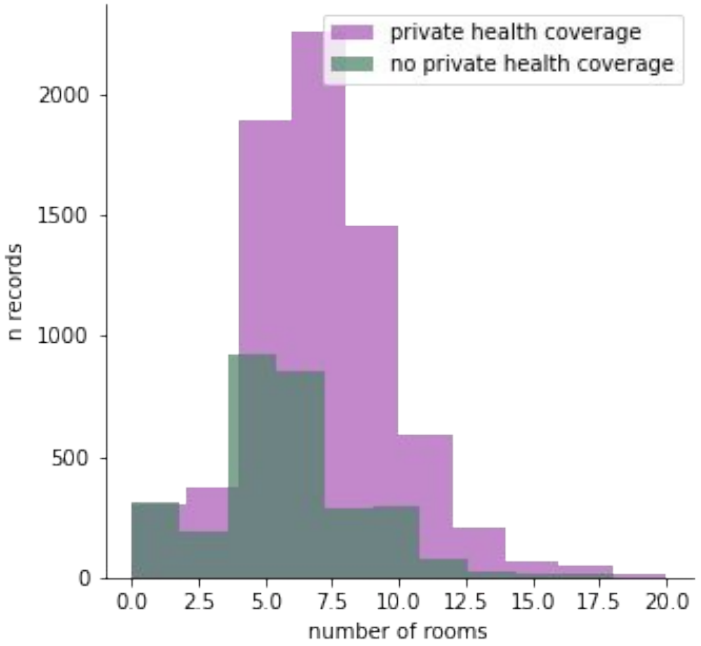
information may be available in regional datasets (organized by zip code, for example)



recommendations based on model features

identify homes with > 8 rooms

information may be available in regional real estate or apartment listing datasets



recommendations based on model features

identify individuals with active internet use

may be available from online sources sharing info collected on websites