Predicting private health coverage in the US

Background

Health insurance covers a person's medical expenses. In the United States, individuals commonly acquire health insurance through public or private options. Most individuals receive private health insurance coverage as a benefit through their employer or labor union[1]. Others opt to purchase private health coverage directly from the insurance company or a government marketplace. In addition to these private options, public health insurance programs, such as Medicare and Medicaid, provide coverage to individuals 65 and older or to individuals with low income, respectively.

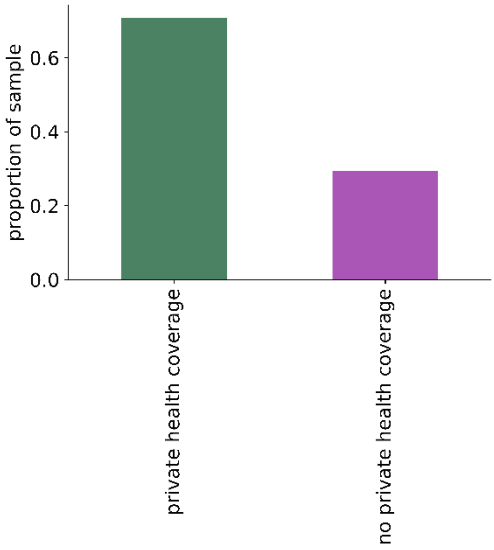


Figure 1. Proportion of individuals with private health coverage

This project aims to identify factors common to individuals in the private health insurance marketplace. Here, a classification model will be developed using demographic data available from the US Census Bureau to predict whether an individual record corresponds to someone with private health coverage. While this dataset is de-identified, the factors common to those may reveal demographic information (features) that are common to those in the private insurance market space. This information may help to identify candidates for targeted advertisements for private health insurance coverage.

Approach

The US Census Bureau collects a wide range of demographic data, related to employment, income, education, ancestry, household information as well as health care coverage, from 1% of the population each year as a part of a program called the American Community Survey[2]. Each record in the dataset represents one individual surveyed, and all records are de-identified. The data are publicly available for download or retrieval from the US Census Bureau API [3]. This rich dataset offers the opportunity to identify demographic information common to those in the private health insurance marketspace.

Python code, written in Jupyter notebooks and Python scripts, was used to retrieve and analyze data from the American Community Survey. Raw and processed data were saved in csv or pickle files.

First, the US Census Bureau API was called to retrieve the American Community Survey data from 2019 (note 2020 data is available for download but not through the API). Because a limited amount of data can be retrieved in a single call, data was

collected for a subsample of individuals (n=200) from each state (and DC). Records from subsamples from each state and DC were eventually joined in a Pandas data frame for further analysis.

Information about variables included in the American Community Survey is available on the API. The target variable, private health insurance coverage (PRICOV), was named, and all other variables related to insurance were dropped to eliminate information conceptually similar to the variable the project aims to predict.

Data dimensions were reduced by eliminating variables with very little variability (i.e., may not hold useful information for distinguishing the target variable), or with many (> 33%) missing values. Following application of these variable filters, over 250 variables remained. To identify which variables are most relevant to the target variable (private health coverage), the predictive power score—a measure that reflects how well one variable informs another variable based on a decision tree—was computed for each candidate feature in the dataset. The distribution of predictive power scores was between 0 and ~0.22. To determine which variables to include as predictors for model development, K-means clustering was performed with predictive power scores as the single feature. Based on silhouette scores, the best value of k of those tested (2-4) was 4. The maximum predictive power score of the bottom two clusters was used to eliminate variables with low predictive power scores. This resulted in 77 features—71 categorical and 6 numeric.

Prior to training classification models, all categorical variables were one-hot encoded. Data were then split into training (70% of records) and test sets, stratifying based on the target variable. Then, all numeric variables in the training set were standardized, and this transformation was applied to the test set.

Five classification models were trained and evaluated--logistic regression, K-nearest neighbors (KNN), random forest, an extreme gradient boosting (XGBoost), and a support vector machine (SVM) model—prior to choosing a final model that best fits the goals of the project. Prior to model fitting, random under-sampling was performed to attempt to counteract the imbalanced dataset for all classification models except for XGBoost. For the XGBoost model, a hyperparameter was included that weights errors for records in the minority class differently in an effort to counteract the imbalance. For those models particularly sensitive to data dimensions—logistic regression, KNN, and SVM models—the first two principal components of the numeric features replaced the six numeric features (net loss of four predictors). For each classification model, hyperparameters were tuned through a grid search with five folds. Training data was fit, and the fit model was used to predict the target variable from predictors in the testing set. Several evaluation metrics were computed including accuracy, F1 score, precision score, and recall score.

Summary

Congruent with prior reports[1], the majority of individuals in this sub-sample have private health coverage (**Figure 1**). As part of the feature selection process, predictive power scores were computed to determine how well the private health insurance coverage could be predicted from a single feature. The majority of features with the highest predictive power scores, and subsequently used in modelling, were categorical. For example, access to the internet and yearly food stamp recipiency were two examples of features with high predictive power. Indeed, visualizing the proportions of individuals without access to the internet and receiving food stamps, grouped by whether they have private health insurance or not, revealed striking differences between groups (**Figure 2**).

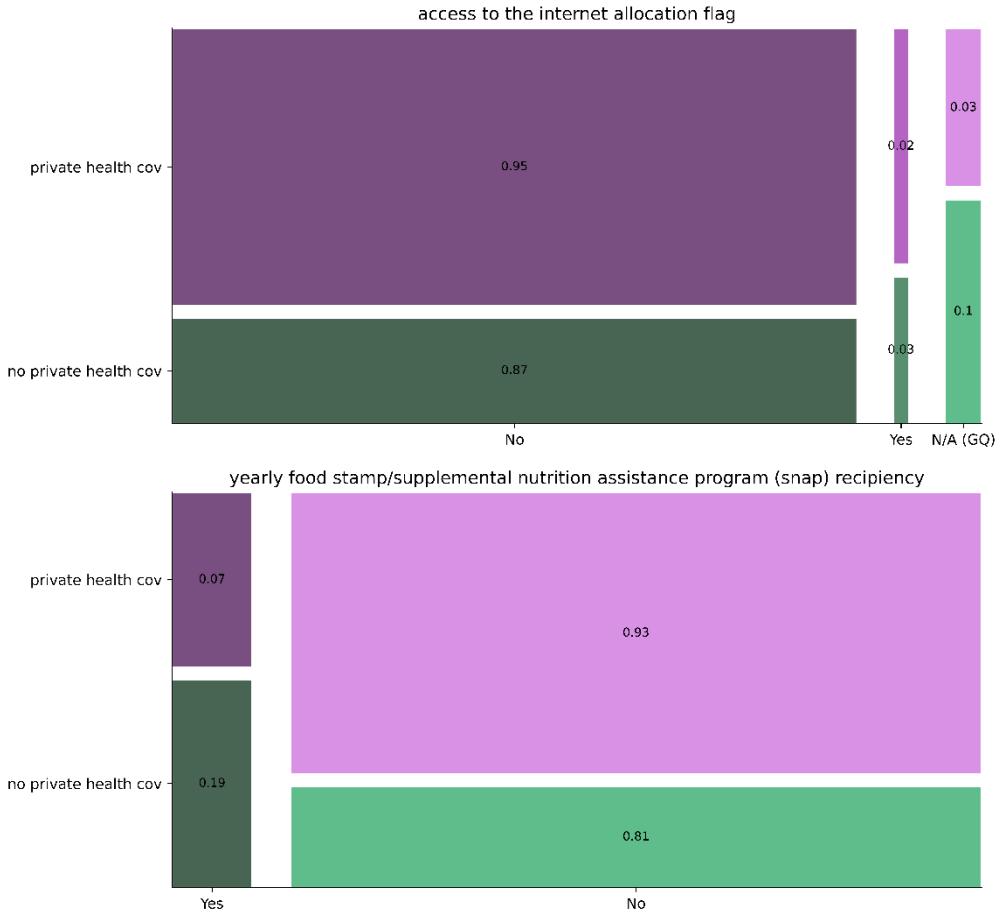


Figure 2. Proportion of individuals with and without access to internet (top) and receiving food stamps, grouped by whether they have private health care coverage.

Several evaluation metrics were computed to assess each of the five classification models. This project assumes marketing funds are limited, and that the target demographic is those with private health coverage already as this demographic may be

in a position to switch to other private healthcare plans. Given spending limits, the precision score was prioritized for model comparison because this metric should maximize identifying individuals with private health coverage while minimizing the number of false positives (incorrectly predicting someone has private health coverage when they do not). The model with the highest precision score was the XGBoost model (0.869; **Figure 3**), followed closely by the Random Forest model (0.865). The overall accuracy was also greatest for the XGBoost model (0.739). All models performed better than a dummy classifier on every metric.

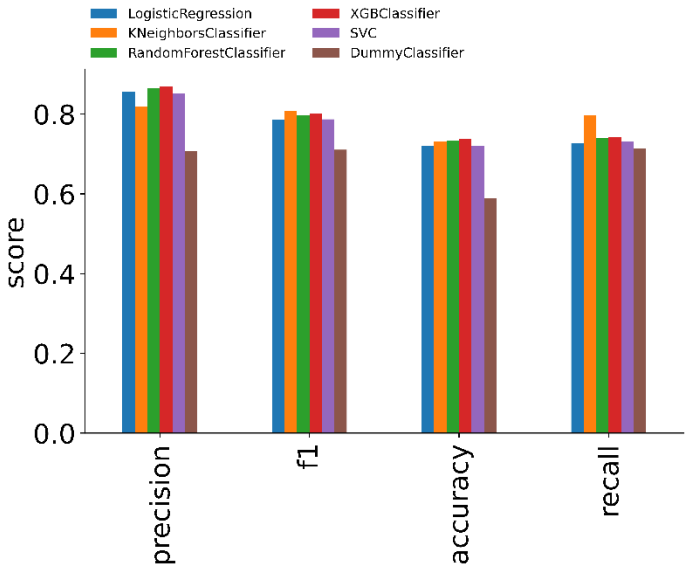


Figure 3. Classification models performance comparison

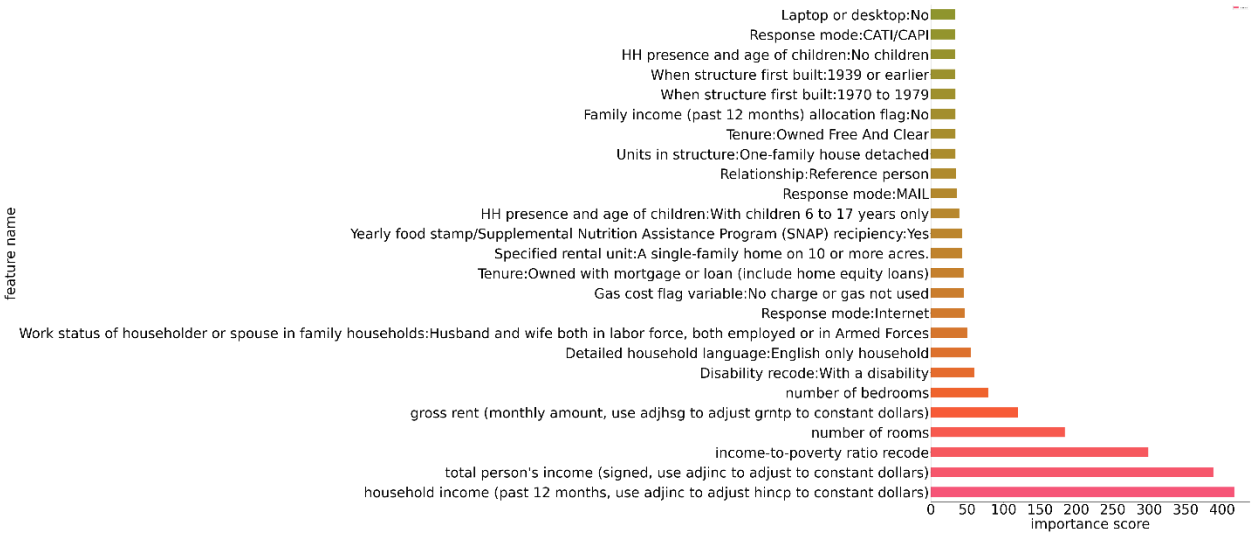


Figure 4. Factors important for predicting coverage with XGBoost model

A major goal of this project was to determine the factors that predict whether someone has private health coverage. One benefit of the XGBoost model is that the importance of the predictors can be assessed by looking at how many times that predictor is used to distinguish the two classes or split the data into those with private health coverage and those without. This metric revealed that predictors related to income (household income, individual income, income-to-poverty ratio) were the three most important factors for solving this classification problem (**Figure 4**). Other important factors included those related to housing (number of rooms, bedrooms, cost of rent).

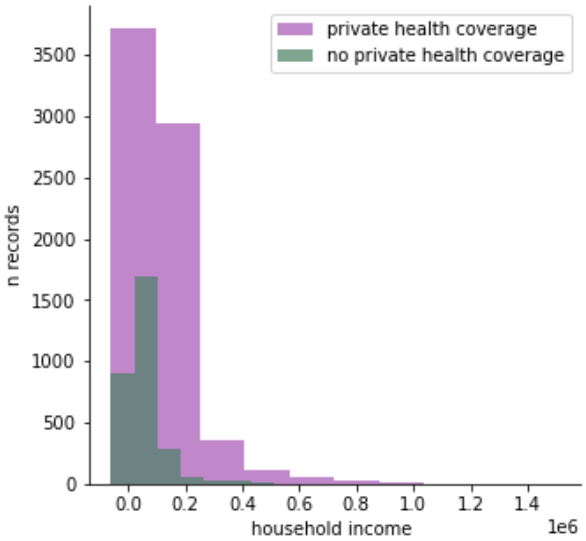
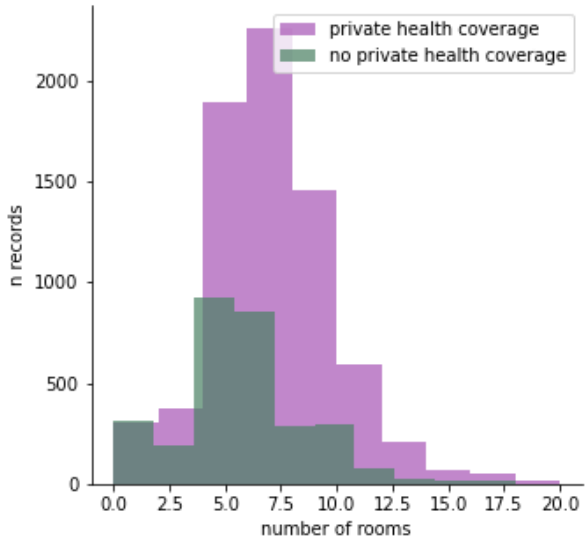


Figure 5. Household income, the most important feature for the XGBoost classification model, grouped by private health coverage.

If a private health insurance company aimed to advertise to individuals with private health coverage in an attempt to acquire new subscribers, they could target individuals known to be in households making over $100,000 per year (**Figure 5**).



Another important feature was for this classification problem was the number of rooms in the house/residence. Grouping the records by private health coverage and visualizing the number of rooms reveals that, while there is good overlap between the groups, residences with 8 or more rooms tend to be inhabited by individuals with private health coverage (**Figure 6**). Lastly, the way the American Community Survey responses were submitted was also a strong predictor for this classification mode (**Figure 7**).

Figure 6. Number of rooms in residence, an important feature for the XGBoost classification model, grouped by private health coverage.

To put this insights from the model into practice, a marketing division may seek demographic information by region, or zip code, to pinpoint areas where there are more likely to be higher household incomes, larger homes, and larger rents. The response

mode from the survey (computer assisted interviews, internet, and mail) also suggest that those with access to the internet (and actively using it) may be good targets for private health coverage marketing as well.

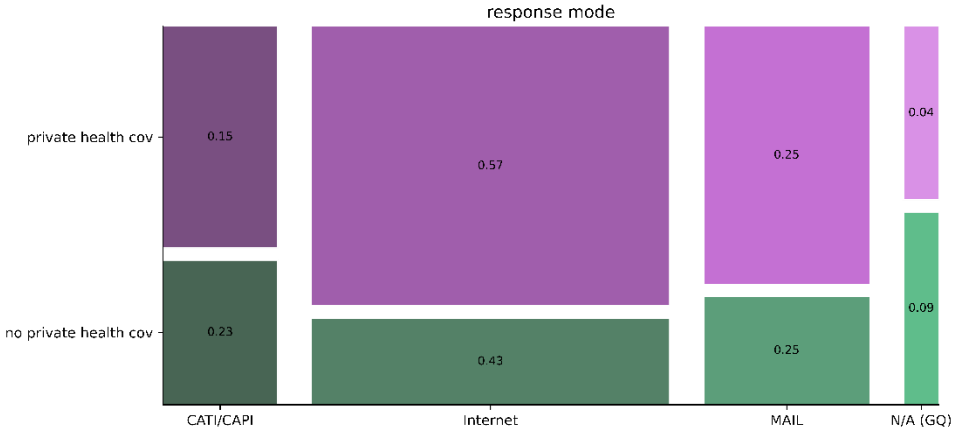


Figure 7. Survey response, an important feature for the XGBoost classification model, grouped by private health coverage may provide insight into other factors (preference for internet or phone) useful for marketing.

References

1. Keisler-Starkey & Bunch (2021). Health Insurance Coverage in the United States: 2020. https://www.census.gov/content/dam/Census/library/publications/2021/demo/p60-274.pdf

2. American Community Survey (ACS) (census.gov). https://www.census.gov/programs-surveys/acs/

3. American Community Survey Data via API (census.gov). https://www.census.gov/programs-surveys