HUMANA-MAYS HEALTHCARE ANALYTICS 2020 CASE COMPETITION

Transportation Challenge Prediction Project

INTRODUCTION

Compared to the extensive awareness of pathological determinants of health, the amount of public knowledge on Social Determinants of Health (SDOH) is insufficient, if not meager. SDOH are conditions in the places where people live, learn, and play that affect a wide range of health and quality-of life-risks and outcomes. [1] Although these determinants may not be observed directly, their pervasiveness in people's lives should not be overlooked, given that 60% of what creates health has to do with the interplay between socio-economic and community environments and lifestyle behaviors.[2] Among these conditions, transportation challenges undoubtedly play an important role. In fact, research shows that there are 3.6 million Americans who do not obtain medical care due to transportation barriers .[3] From the Medicare-provider perspective, such challenges also have a negative impact given that it incurs a $150 million cost of missed appointment to the industry.[4]

If transportation challenges have been a dilemma that patients could handle using ad-hoc methods including public transportation or ride-share services over the years, the recent COVID-19 pandemic assuredly worsened the problem. With limited public transportation means available and a fatal virus spreading across the world, patients are being forced to make a difficult decision between risking infection going to the medical appointment or staying home but leaving symptoms untreated. Thus, we face an unprecedented urgency to quickly and accurately identify people who are at risks of transportation challenges. And then we need to develop a feasible plan to remove the social determinant barriers for high risk people, so that they can receive quality healthcare.

BUSINESS OPPORTUNITY IN TRYING TIMES

Arising from these Medicare crises are business challenges and opportunities. As has been discussed above, the population being influenced by transportation challenges amounts to 3.6 million in the U.S., a significant size worth paying attention to. For this group of people, it is more likely to have disruptions of patient care and provider-patient relationships, delayed care and increased emergency department visits.[5] It then becomes vital from a public health point of view to address this issue.

[1] https://www.cdc.gov/socialdeterminants/index.htm

[2] https://about.kaiserpermanente.org/community-health/news/making-a-down-payment-on-health-kaiser-perm
anente-invests-in-cre

[3] 1Health Research & Educational Trust. (2017, November). Social determinants of health series: Transportation and the role of hospitals. Chicago, IL: Health Research & Educational Trust. Accessed at www.aha.org/transportation

[4] Sviokla, J., Schroeder, B. & Weakland, T. (2010). How behavioral economics can help cure the healthcare crisis. Harvard Business Review. Retrieved from https://hbr.org/2010/03/how-behavioral-economics-can-h

[5] Syed, S. T., Gerber, B. S. & Sharp, L. K. (2013). Traveling towards disease: Transportation barriers to healthcare access. Journal of Community Health, 38(5): 976-993.

Besides, from the industry practitioner's perspective, catering to the needs of those having transportation challenges is a good strategy to save millions of dollars from missed appointments. Statistics show that missed appointments and the resulting delays in care cost the health system $150 billion each year in the U.S.[6] During such difficult times where cash flow becomes stringent because of economic downturn under a global pandemic, coming up with cost-effective strategies in operations becomes imperative for business stakeholders, particularly those in the Medicare domain. In addition, under such difficult times, healthcare providers who are willing to dig deep into the issue and address the challenge can establish a trust-worthy, member-friendly brand image. This can help build a solid bond between healthcare providers and Medicare members as well as create positive connections with potential clients.

RESEARCH QUESTION

This analysis aims at using statistical modeling and machine learning algorithms to help identify Medicare members who are most likely to experience transportation challenges. Based on the major findings, this model also proposes viable solutions for families, communities, Medicare-providers, and policy makers to remove these barriers for patients to access care and achieve their best health.

BACKGROUND

Transportation screening question came from the Accountable Health Communities – Health Related Social Needs Screening Tool. Medicare members were asked to provide a ‘Yes’ or ‘No’ answer to the question: “In the past 12 months, has a lack of reliable transportation kept you from medical appointments, meetings, work or from getting things needed for daily living?” It is likely that members struggling with transportation challenges are not homogenous and hence there are perhaps different solutions for different segments of members.

MAJOR FINDINGS

- Clinical-related conditions, credit balance, pharmacy/prescription claims, and number of times physician office visits are the four most informative factors affecting transportation challenges.
- It is important for communities and the government to allocate resources for low-income groups to reduce the impact of transportation issues on health outcomes.

[6] Health Research & Educational Trust. (2017, June). Social determinants of health series: Food insecurity and the role of hospitals. Chicago, IL

- Government and policy makers should provide support for high risk community members and develop better infrastructure for more reliable public transportation.
- Hospitals should also closely monitor patients who are more likely to have transportation issues and pay close attention to their health outcome.

DATA PREPARATION

Data Source (800+ features)
Raw training data and a holdout set are provided by Humana. Data come from various sources document members' information, including:

- Medical claims features
- Pharmacy claims features
- Lab claims features
- Demographic / Consumer data
- Credit data features
- Clinical Condition related features
- CMS Member Data elements
- Other features

Each member has a binary flag to indicate transportation challenges.

Training data have 69, 572 observations for 826 variables and the holdout set has 17, 681 observations for 825 variables. Among those, 59, 375 of the observations reported not having transportation issues, while 10, 197 reported having transportation issues.

Analytics Tools

- R Studio for preliminary data inspection
- Interactive Python in Jupyter Notebook
- Pandas and Numpy for feature extraction
- Scikit-learn for feature engineering and model development
- Matplotlib for data visualization
- Amazon Web Services (AWS) for hyperparameter tuning
- Google Cloud Platform (GCP) for Auto ML
- Excel for data exportation

ASSUMPTIONS

In the preliminary research stage, we first hypothesized that it would be most effective to engineer features that are related to clinical-related conditions, educational levels, and household composition to examine their potential influences on transportation challenges.

FEATURE ENGINEERING

1. **Variable Selection**
   After examining the data types, correlations among each feature group, and the strength of linear relationship of each feature with the response variable, we removed variables with redundant information and collapsed certain variables from multiple groups into fewer levels to reduce overfitting and to increase comprehensibility. Below are the four major subsets of variables selected after preliminary screening:

- Clinical-condition related features (BETOS codes, MCC categories, etc.).
- Credit data features (credit balance of various account types).
- Medical/Pharmacy/Lab claim features (CMS revenue, risk adjustment payment rate, etc.).
- Demographic/Consumer data and other features (gender, age, geographic information, preferred language, smoke status, household composition, indicators of disability and low income, etc.).

2. **Missing Value Handling**
   The training set contains a total of 21, 725 observations with at least one missing value. Since the number of observations that contain missing values is relatively large, it would not be ideal to drop these observations. We imputed the missing values by implementing the K-nearest neighbors (KNN) clustering technique with K = 2.

3. **Dimension Reduction**
   Due to high dimensionality of the dataset, we performed PCA on each of the first three subsets of variables specified in the Variable Selection section above to reduce dimensions and avoid overfitting.

   Applying the PCA algorithm, we reduced:

- Clinical condition-related dataset from 224 features to 80 principal components for each Medicare member, capturing 90.0% of the variance within the dataset.
- Credit dataset from 20 features to 5 principal components, capturing 99.7% of the variance within the dataset.

- Medical/pharmacy/lab claims dataset from 13 features to 4 principal components, capturing 99.9% of the variance within the dataset.

  Upon completing the imputation of missing values and feature engineering, along with 40 variables from demographic/consumer related features, our dataset was reduced to 129 features in total.

MODEL EXPLORATION

Our goal is to find the best model with the most important features to predict the probability of each Medicare member experiencing transportation challenges. Therefore, we aimed to find a model with the highest AUC score, as it provides an aggregate measure of performance across all possible classification thresholds.

As we began to explore potential models, we divided the data from the reduced dataset (129 variables) into three subsets:

- Training set: consisting of 60% of the observations derived from the reduced dataset.
- Test set: consisting of 20% of the observations derived from the reduced dataset.
- Validation set: consisting of 20% of the observations derived from the reduced dataset.

We used the training set to train models, the test set to tune model hyperparameters, and validation set to estimate AUC scores.

Logistic Regression

Since having transportation issues or not is a classification problem, we first implemented a Logistic Regression. The model reached an AUC score of 0.54. Although we did not end up incorporating this model into our final solution, this model served as a baseline as we attempted to explore more complex solutions for better results.

Random Forest

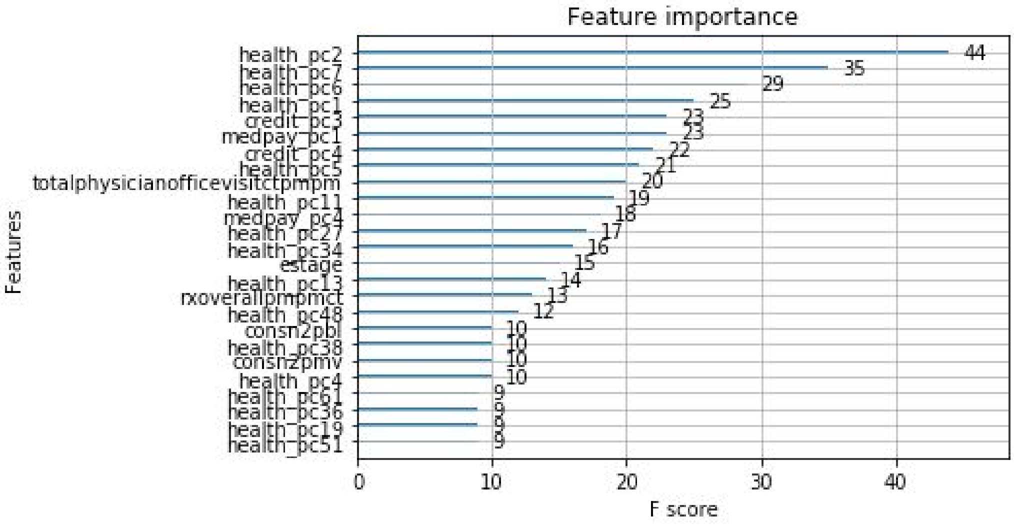
Moving on to the next step of model selection, we used a random forest classifier in the analysis, which is a meta estimator that uses averaging from various samples of the dataset to improve predictive accuracy and control over-fitting. On the test set, the model reached an AUC score of 0.73.

XGBoost

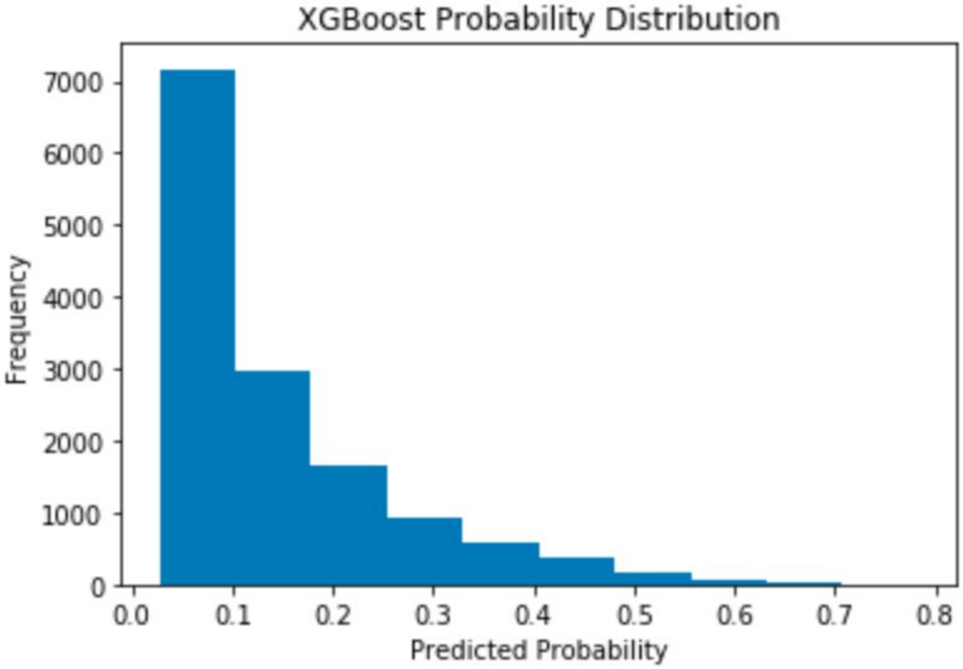
The next model we implemented was the Extreme Gradient Boosting model (XGBoost). It is an implementation of gradient boosted decision trees that is optimized for speed and performance. The XGBoost model uses regularization parameters to avoid overfitting. When training the model, we took advantage of the AWS Sagemaker hyperparameter tuner module to find the best hyperparameters by setting the objective to maximize the AUC score. The hyperparameters that resulted in the highest performing model are shown in the table.

Hyperparameter	Metric
alpha	69.2
eta	0.2
min_child_weight	3.3
subsample	0.9
num_round	50

Using 60% of data as the train set, 20% as the validation set, and 20% as the test set, the AUC score for the test set is 0.74. This was the best performing model among all models. The plot below shows the feature importance for 25 most important variables. It is interesting to note that health-related, finance-related, and medical payment-related variables, total physician office visit, prescription variables are important in predicting the probability of transportation challenges. This finding is consistent across all models.



The distribution of the probabilities is shown in the graph below. It shows that most of the observations are less likely to experience transportation challenges.

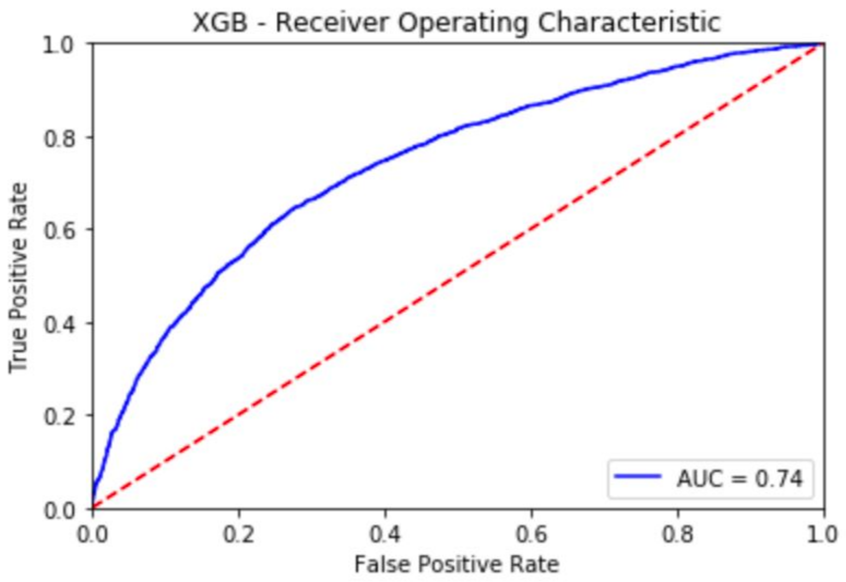


FINAL MODEL SELECTION

A summary of AUC scores for all three models is shown in the table.

Model	AUC
Logistic Regression	0.54
Random Forest	0.73
XGBoost	0.74

The XGBoost model outperformed other models in terms of AUC score. The ROC-AUC plot is shown below. We decided to use the XGBoost model as our final model to make predictions for the holdout set.



FINAL ANALYSIS AND ACTIONABLE INSIGHTS

According to the most informative features selected from the XGBoost model above, principal component features regarding clinical-related conditions (health_pc2, health_pc7, health_pc6, health_pc1) have the most predictive power on transportation issues, followed by a principal component feature of credit balance (credit_pc3), medical claims (medpay_pc1), physician office visits, and pharmacy claims (rxoverallpmpmct). Race, household composition, home-owner status, and other socio-economic status also contribute to the probability of having transportation issues. The F-score serves as a measure of how informative each feature is for the dataset, representing the ratio between the explained and the unexplained variance.

Transportation challenges as a SDOH interplays with a lot of other determinants, making it difficult to predict and no less challenging to address. Also, it is likely that members struggling with transportation challenges are not homogenous, there are different solutions for different subgroups of members.

Based on the findings above, we divide actionable steps into screening, accommodation, and feedback stages. We advise that in order to improve the health of patients' communities, different means and strategies should be deployed by different responsible parties, including but not limited to, communities, hospitals, and policy makers. We also call for a better collaboration and information-sharing between service-providers and the local community to form a support network.

1. **General Screening**

a. When collecting information from members, it is necessary to put more emphasis on getting detailed financial, health-related, and claim-count related information, so that the model will have higher predictive power as identifying patients with transportation challenges.

b. The trained model can be easily scaled up with a serverless AI engineering pipeline using a Flask app as soon as members' information is encoded into the system. After deploying the model for real-time prediction, people who are at greater risks of experiencing transportation challenges can be identified as early as possible.

c. Information on transportation challenges can be shared with community level workers, who can pay more attention to people who are at higher risks of encountering transportation challenges and access to healthcare.

2. **Accommodations and Adjustments**

a. Based on the prediction given by the model for each individual member, those identified as being more likely to have transportation challenges can be categorized into a specific database where they are contacted for available accommodations.

b. The prediction information can also be used for care-providers so that they can know which patients may be more likely to miss appointment / prescription refill so that they can arrange for more flexible schedules and then less likely to disrupt patient-provider relationships.

c. Approaches for different subgroups:

i. Socio-economic status:

1. Members identified as having lower socio-economic status can be supported with free transportation services and volunteers. These services may provide a flexible window period for them to use.
2. Members identified as having higher socio-economic status can be assisted by volunteers or social workers to obtain easier access to transportation services.

ii. Rural-urban residents:

1. Providers should stay in touch with members in the rural areas who are more likely to have transportation challenges. Those members in the same community can be grouped together and assistance can be provided on a small-community basis.
2. In the urban areas, a more reliable and safer public transportation system, such as bus and subways, is needed to reduce the likelihood of transportation issues. Similar to rural residents, fostering a small, close-knit community can provide support and assistance in a more timely manner for urban members.

iii. Health condition

1. For those with complications or multiple health conditions, having transportation challenges can be fatal. It is necessary for hospitals and providers to closely collaborate and monitor their activities.
2. For those with less severe health conditions, forms of assistance mentioned above can be provided independently or as a combination based on the specific needs of patients.

3. **Feedback and Optimization**
    a. Track members who are at higher risks of having transportation issues and getting feedback every 3 months. Update their information and whether or not they still have the challenge.
    b. Collect information on patients who have frequently missed appointment / prescription refills, and include this information in the model to improve performance. These patients should be added to the support network for further assistance.

STRATEGIES AND PLANS

- Understanding and examining the impact of social determinantes, especially transportation access, on public health.
- Assessing and evaluating public transportation on health benefits.
- Identify the benefits of minimizing transportation challenges, such as fewer missed appointments and delayed care.
- Developing strategies and allocate resources for safe and reliable transportation options for low-income groups in urban and suburban areas.
- Supporting programs, policies, and technology development to improve transportation access and efficiency.
- Developing other online services to streamline virtual medical supports, such as MyChart and other mobile clinics, to reduce the chance of accessing in-person medical services.

FUTURE CONSIDERATIONS

In conclusion, since the consequences of having transportation challenges may lead to poor management of chronic disease, and thus undesirable health outcomes, it is important that hospitals, communities, and policy makers take actions to reduce the transportation barriers to healthcare access. Future research needs to focus on:

- Identifying and understanding patients’ transportation needs.
- Spotting the different aspects and areas where transportation is most pronounced that affect health access.
- Measuring the impact of transportation barriers against policy changes.

APPENDIX

Figure 1. ROC for Logistic Regression

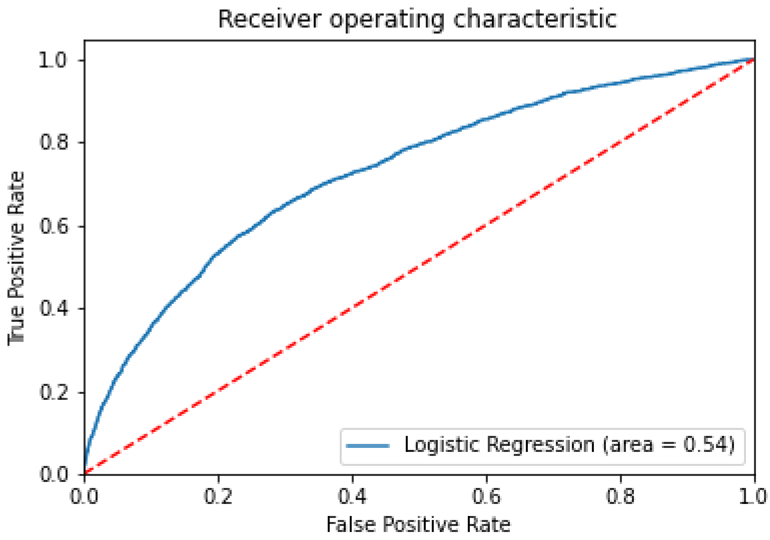


Figure 2. ROC for Random Forest

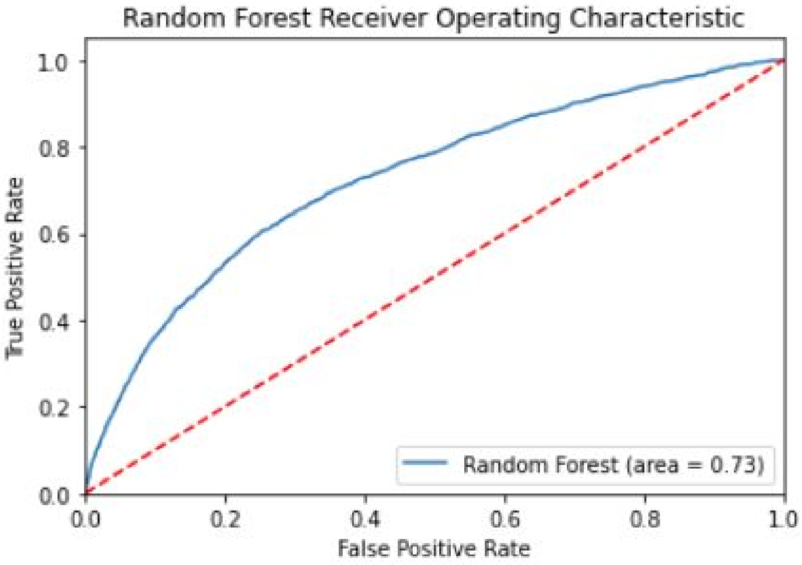


Figure 3. XGBoost Tree Plot - Number of Trees = 2

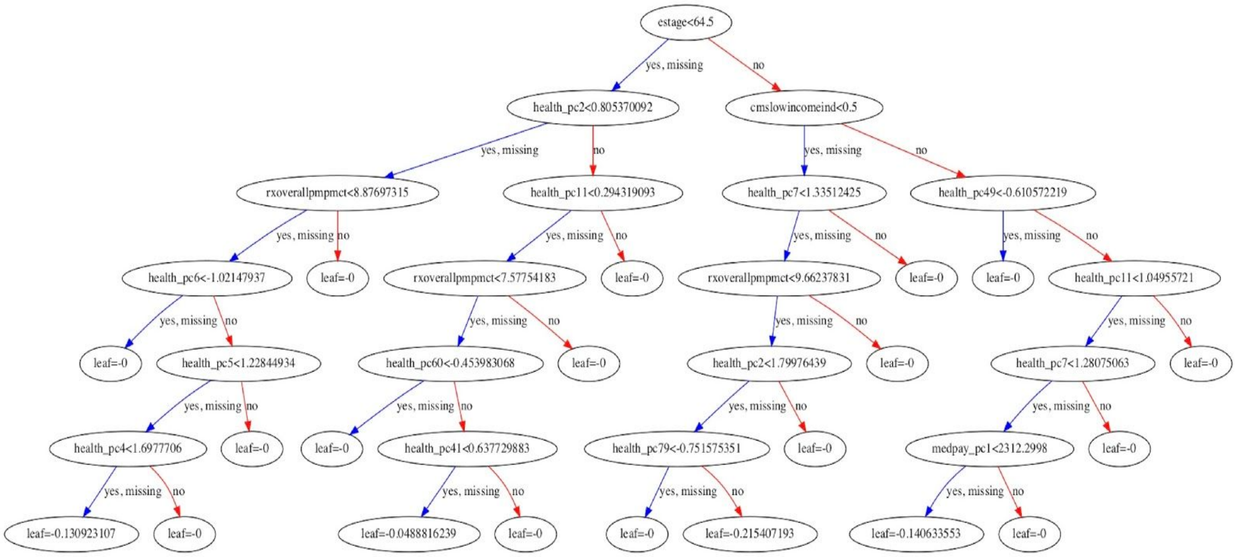
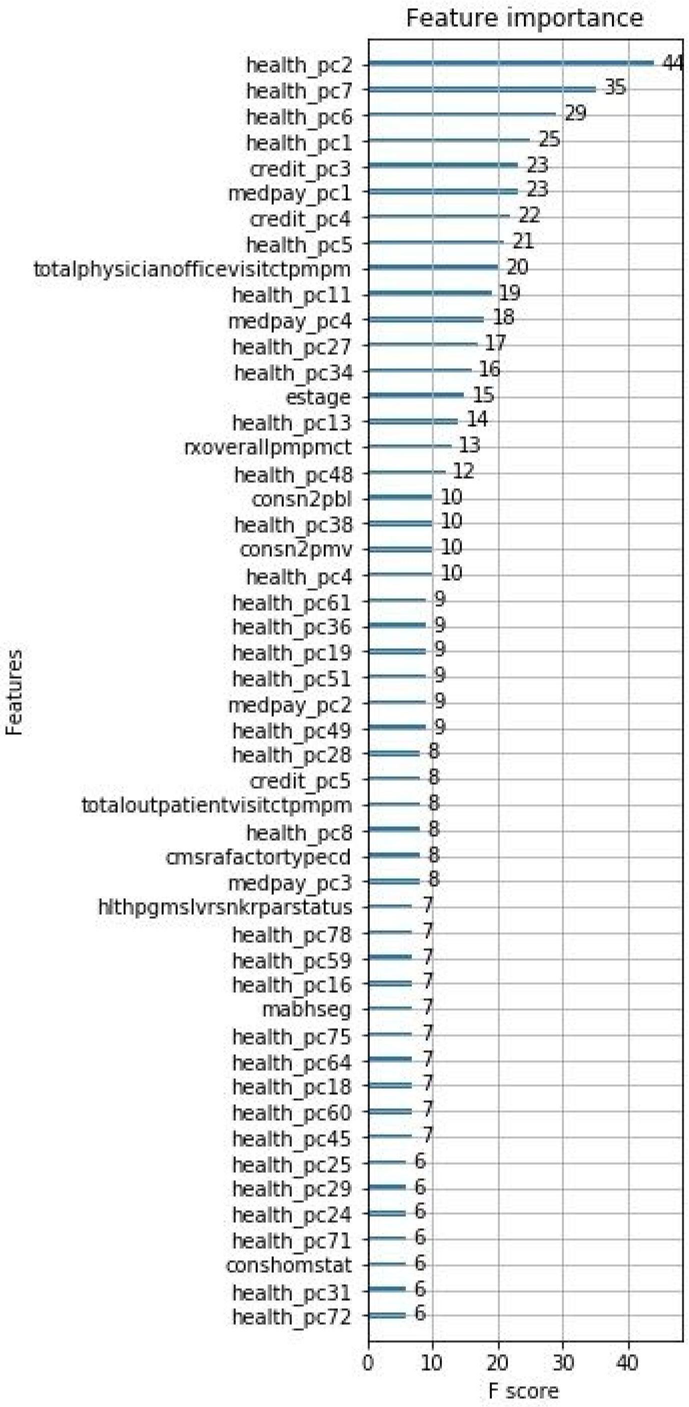


Figure 4. XGBoost Top 50 Importance Features



Please check out the complete python code in this GitHub repository.