# Humana-Mays Healthcare Analytics 2022 Case Competition

Housing Insecurity Issues: Prediction and Segmentation Analysis

October 16th, 2022

# 1 EXECUTIVE SUMMARY

## 1.1 Study proposal

In the U.S, the housing insecurity issues are increasingly severe. Housing insecurity encompasses a number of challenges, such as having trouble paying rent, overcrowding, moving frequently, or spending the bulk of household income on housing. These experiences may negatively affect physical health and make it harder to access health care. As a leading healthcare company offering a wide array of insurance products and health and wellness services, Humana is dedicated to addressing members' housing insecurity problems. This study focuses on helping Humana achieve its bold goal by applying big data analysis and machine learning methods. Our objective is to identify members who are most likely to experience housing insecurity issues and offer corresponding recommendations.

## 1.2 Modeling

In order to achieve the best performance of modeling, we carried out comprehensive studies in understanding the business issue we need to fix and all the features in the dataset. First we built a predictive model to identify members who are the most likely to experience housing insecurity issues. We chose Gini Index, random forest and XGBoost to do feature selection based on three models' intersection, developing a better understanding of the most important features included in our model. Then we applied Random Forest, Gradient Boosting Decision Tree, LightGBM and XGBoost, along with parameter tuning to do preliminary prediction and compared their performances and corresponding AUC. Finally we got the best performance with an AUC of 0.761 with XGBoost. The further analysis and recommendations regarding improvement of housing security is based on the features we derived.

## 1.3 Recommendation

We identified key drivers of housing insecurity issues and developed scalable business solutions for specific segments of Humana members based on this model. We put forward targeted and personalized recommendations, including providing financial assistance, offering convenient home maintenance and repair, improving living environment and housing quality, establishing health management systems and providing house testing and disinfecting. These recommendations are designed to improve the overall health outcomes of Humana members by addressing these housing insecurity issues and to increase Humana's effectiveness and reputation.

# 2 CASE BACKGROUND

## 2.1 Context

Housing insecurity is one of the most important components of health-related social needs, which are the immediate health-harming conditions affecting a specific individual. Housing insecurity is defined as lack of access to quality and safe housing, including lack of safe, affordable, and stable housing. Housing insecurity is associated with health problems related to both physical and mental health. Housing instability such as frequent move or evictions may lead individuals to injury, disease, mental illness and behavioral health issues. According to a 2018 AARP survey, three out of four adults age 50 and older want to stay in their homes and communities as they age, while "universal design" elements—such as no-step entries, extra-wide hallways and doors to accommodate walkers and wheelchairs, and lever-style door and faucet handles— can help make homes safer for seniors, only 57% of existing homes have more than one of these features. The cost of making necessary home modifications may be too burdensome for many, forcing individuals to either remain in unsafe living environments or move to nursing homes or long-term care facilities[1]. Based on the research of the National Low Income Housing Coalition, 34.6% of U.S. households were cost-burdened, while 16.3% were severely cost-burdened, paying more than 50% of their income for housing[2]. Not only does this burden increase the chances of housing instability but it often means families struggle to afford basic needs like food and medical care.

## 2.2 Problem statement

Humana is a leading healthcare company that offers a wide array of insurance products and health and wellness services. It serves around 17.1 million members nationwide. The purpose of this analysis is to help Humana to identify its members who are most likely to be struggling with housing insecurity problems and provide potential recommendations and solutions to improve members' living quality and health conditions. To achieve this goal, we first applied a classification model to predict which members are most likely to experience housing insecurity issues based on the provided data. Then we identified the most important features affecting housing insecurity, categorized the members into five major groups and proposed potential solutions to address their housing insecurity problems.

---

[1] Joanne Binette, Kerri Vasold, AARP Research, August 2018, Revised July 2019; 2018 Home and Community Preferences: A National Survey of Adults Ages 18-Plus

[2] National Low Income Housing Coalition, "Out of Reach 2021: The High Cost of Housing report"

# 3 DATA ANALYSIS

## 3.1 Dataset description

Humana provided the following two data set of members healthcare information – the training set used to train the model and the holdout data set to be predicted:

- training data: 48,300 records by 881 variables, with a response column hi_flag. 'hi_flag = 1' means that members have housing insecurity problems, accounting for 4% of all members, and 96% of members reporting housing insecurity problems.
- holdout data: 12,220 records by 880 variables.

Among all variables we have, we preliminarily divide our features in below 4 groups: medical claims, pharmacy claims, demographics information, and other important information about members' living conditions.

Table 1. Features Categories

| ***Group (# of variables)*** | ***Prefix/Name*** | ***Feature Description*** | ***data type*** | ***Scale*** |
|---|---|---|---|---|
| **Medical claims and condition features(436)** | cmsd1 | Claim count by CMS diagnosis code categories | numerical(float) | Personal level |
| | cmsd2 | | numerical(float) | Personal level |
| | med | Days since last claim for non-behavioral health claims | numerical (integer) | Personal level |
| | bh | Claim count and cost for behavioral health claims | numerical(float/ double) | Personal level |
| | total | Days since last claim/visits/allowed cost per month for overall claims | numerical(integer/d ouble) | Personal level |
| | cci | Charlson Comorbidity Index and utilization | numerical(float) | Personal level |
| | dcsi | Diabetes Complication and Severity Index score | numerical (integer) | Personal level |
| | rx_[?]_pmpm_ct | Prescriptions based on categories of drugs | numerical(float) | Personal level |
| | rx_[?]_pmpm_cost | | numerical(float) | Personal level |
| | rx_tier_[1,2,3,4]_pmpm_ct | 4 kinds of tier drugs | numerical(float) | Personal level |

| | | | | |
|---|---|---|---|---|
| | rx_hum_[?]_pmpm_ct | Different Drugs under Humana prescription | numerical (float) | Personal level |

| | | | | |
|---|---|---|---|---|
| **Pharmacy Claims Features (234)** | rx_hum_[?]_pmpm_cost | categories | numerical(float) | Personal level |
| | rx_pharmacies_pmpm_ct | # pharmacy | numerical(float) | Personal level |
| | rx_phar_cat_[?]_pmpm_ct | prescriptions in different kinds of pharmacies | numerical(float) | Personal level |
| | rx_perphy_pmpm_ct | # physicians | numerical(float) | Personal level |
| | rx_overall_pmpm_ct | Total prescriptions | numerical(float) | Personal level |
| | rx_overall_pmpm_cost | | numerical(float) | Personal level |
| | rx_days_since_last_script | Days since last prescription in the past one year | numerical(float) | Personal level |
| **Demographics/ CMS/ Consumer Features(29)** | est_age | Age | numerical (integer) | Personal level |
| | sex_cd | Gender | Categorical (string) - > binary | Personal level |
| | cms_race_cd | Race | Categorical(string) | Personal level |
| | cms_disabled_ind | Disability | numerical (integer) -> binary | Personal level |
| | cms_dual_eligible_ind | Dual eligible | numerical (integer) -> binary | Personal level |
| | cms_low_income_ind | Low income subsidy | numerical (integer) -> binary | Personal level |
| | rucc_category | Rural category | Categorical(string) | Regional level |
| | cms_tot_partd_payment_amt | CMS | float | |
| | cms_institutional_ind | | binary | |
| | cms_ra_factor_type_cd | | string | |
| | cms_risk_adj_payment_rate_b_amt | | float | |
| | cms_hospice_ind | | integer | |
| | cms_orig_reas_entitle_cd | | string | |
| | cms_risk_adjustment_factor_a_amt | | float | |
| | cms_ma_risk_score_nbr | | float | |
| | cms_rx_risk_score_nbr | | float | |
| | cms_partd_ra_factor_amt | | float | |
| | cms_ma_plan_ind | | binary | |
| | cms_frailty_ind | | binary | |
| | lang | Other Demographics | Categorical(string) | Personal level |

| | | | | |
|---|---|---|---|---|
| | atlas | | numerical(double) | Regional level |
| **Other Features(180)** | rwjf | physical environments | numerical (float) | Regional level |
| | prov | provider lines | numerical(float) | |
| | rev | descriptions and dollar amounts charged for hospital services | numerical(float) | Personal level |
| | cnt | member interactions | numerical(float) | Personal level |
| | credit | Credit information | numerical(float) | Personal level |
| | cons_homstat | Homeowner Status | Categorical(string) | Regional level |
| | CONS_MOBPLUS | Mail Order Buyer | Categorical(string) | Regional level |
| | cons_lwcm10 | The probability of the individual not exercising at all | float | Regional level |
| | cons_hxmioc | Managing Illness or Condition - Index | integer | Regional level |
| | cons_hxmboh | Managing the Business of Health | integer | Regional level |
| | cons_stlnindx | Student Loan Index | integer | Regional level |
| | cons_ccip | Census Income Percentile | float | Regional level |
| | cons_stlindex | Short Term Loan Index | integer | Regional level |
| | cons_hxmh | Managing Health - Index | integer | Regional level |

## 3.2 Descriptive Statistics

Humana's dataset provides a wealth of membership information, and before the data cleaning, we had a brief understanding of the general situation of members in the Humana database.

- Age Distribution: Humana's database contains 48,300 members' age information, of which the average age of members is 72.06, and the 25% and 75% quantiles are 67 and 77 years old, respectively. Combined with the histogram of age distribution, we can also observe that most of Humana's members are 60-90 years old.

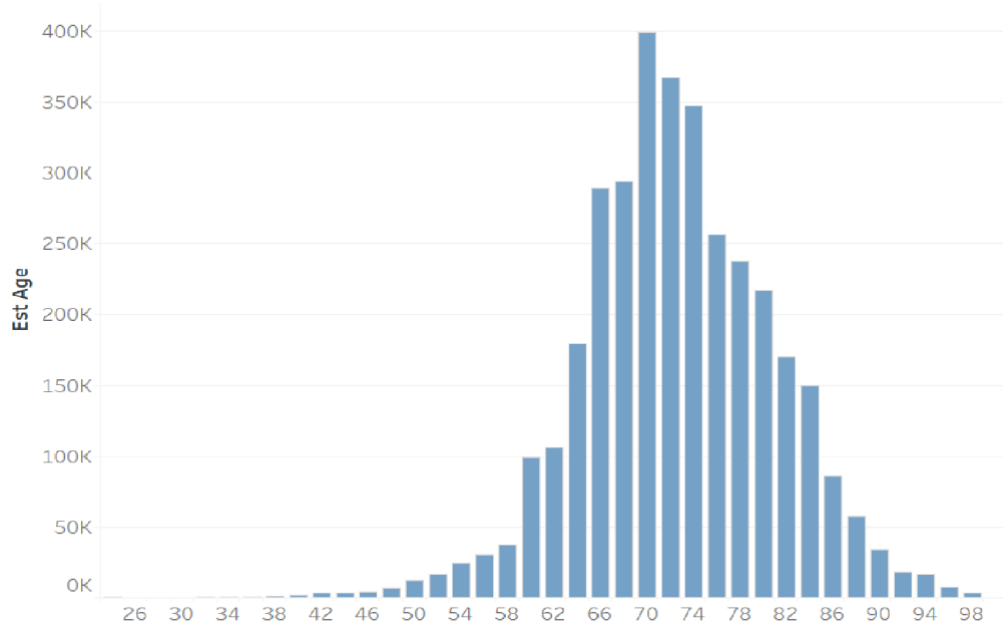


Figure 1. Age Distribution

When we studied the age distribution of these members based on whether there is a situation of housing insecurity, we found some interesting scenarios. Members facing housing insecurity have a lower average age than members who do not have housing insecurity.

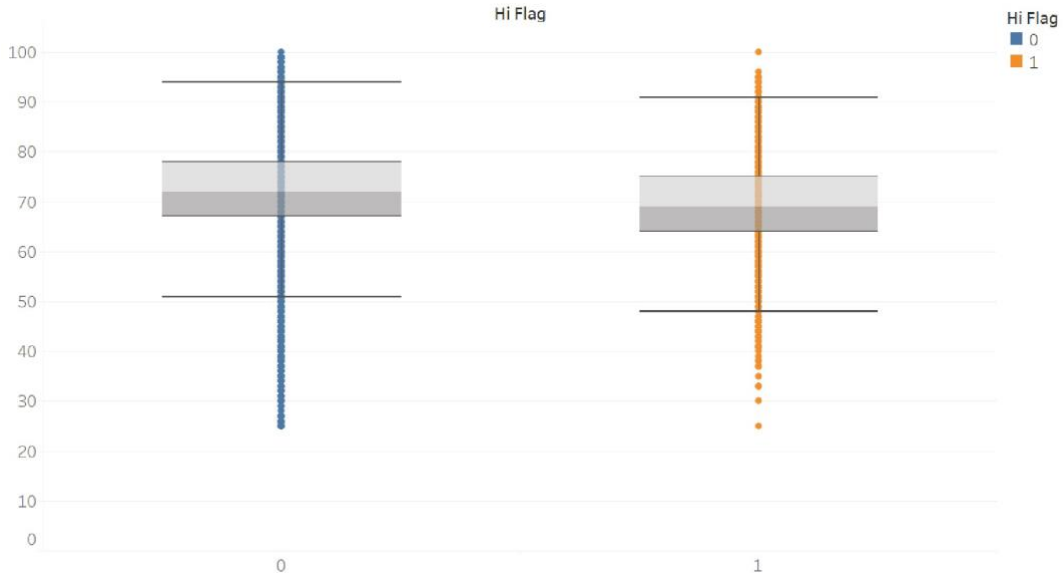


Figure 2. Box Plot by *hi_flag*

- Race: In the database, feature ‘cms_race_id’ records the racial information of members, including Whites (non-Hispanic), Blacks (non-Hispanic), Asian, Hispanic, American Indian or Alaska Native, Others and Unknown. of which the largest proportion is the white group, accounting for 78%. By looking at the racial

distribution of members with housing insecurity, We found that blacks account for 22.19%, which is higher than the proportion of blacks in the total population of 16%.


Figure 3. Pie Chart of Race Distribution

- Member geographic information: The geographical information is mainly divided according to the population size, and is divided into nine categories: category 1 represents counties in metro areas of 1 million population or more; category 2 represents counties in metro areas of 250,000 to 1 million population; category 3 represents counties in metro areas of fewer than 250,000 population; category 4 represents urban population of 20,000 or more, adjacent to a metro area; category 5 represents urban population of 20,000 or more, not adjacent to a metro area; category 6 represents urban population of 2,500 to 19,999, adjacent to a metro area; category 7 represents urban population of 2,500 to 19,999, not adjacent to a metro area; category 8 represents completely rural or less than 2,500 urban population, adjacent to a metro area and category 9 represents completely rural or less than 2,500 urban population, not adjacent to a metro area.

81.3% of members live in metro counties, i.e. fall into categories 1, 2 and 3.

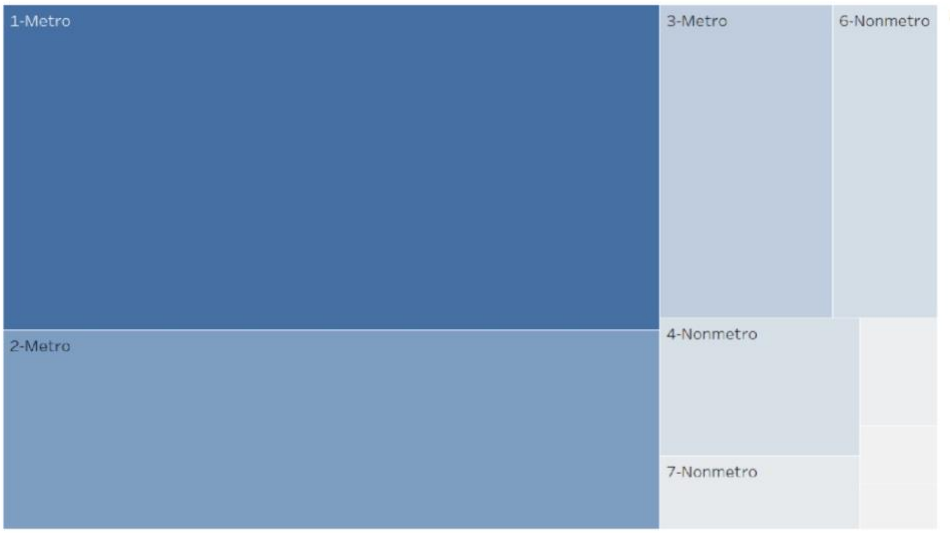

Figure 4. Member Geographics

- Homeowner Status: The homeowner status of members is mainly divided into the following five types:

P = Probable Homeowner
R = Renter
T = Probable Renter
U = Unknown
Y = Homeowner

In the subsequent analysis process, we mainly care about whether members own their own houses. 55.37% of people are houseowner, while 44.63% of members do not own their own house. In addition, we found that considering members who have housing insecurity issues, 30.45% of people are houseowner, which is much lower than 55.37%.

## 3.3 Data cleaning and imputation

### 3.3.1 Data Types Transformation

The first step of our data cleaning is to regulate the data types of all features. We found that some categorical features, such as ‘cms_race_cd’, have inconsistent data types, that is, some data in the feature is of numerical type, and some data is of categorical type, so we first perform type-consistent conversion on all categorical variables.

Next, we convert all missing values to NaN to avoid numerical data being recognized as categorical data in the following data processing process, and also prepare for the next work of missing value imputation.

### 3.3.2 Missing Value Imputation

Before we formally deal with missing values, we first understand the overall situation of missing values in the dataset. We found that there are 47 features with more than 20% missing values. Among them, 30 are 20%-50%, and 17 are more than 50%. Figure 5 is a horizontal histogram of the proportion of missing values (missing rate), sorted in ascending order.

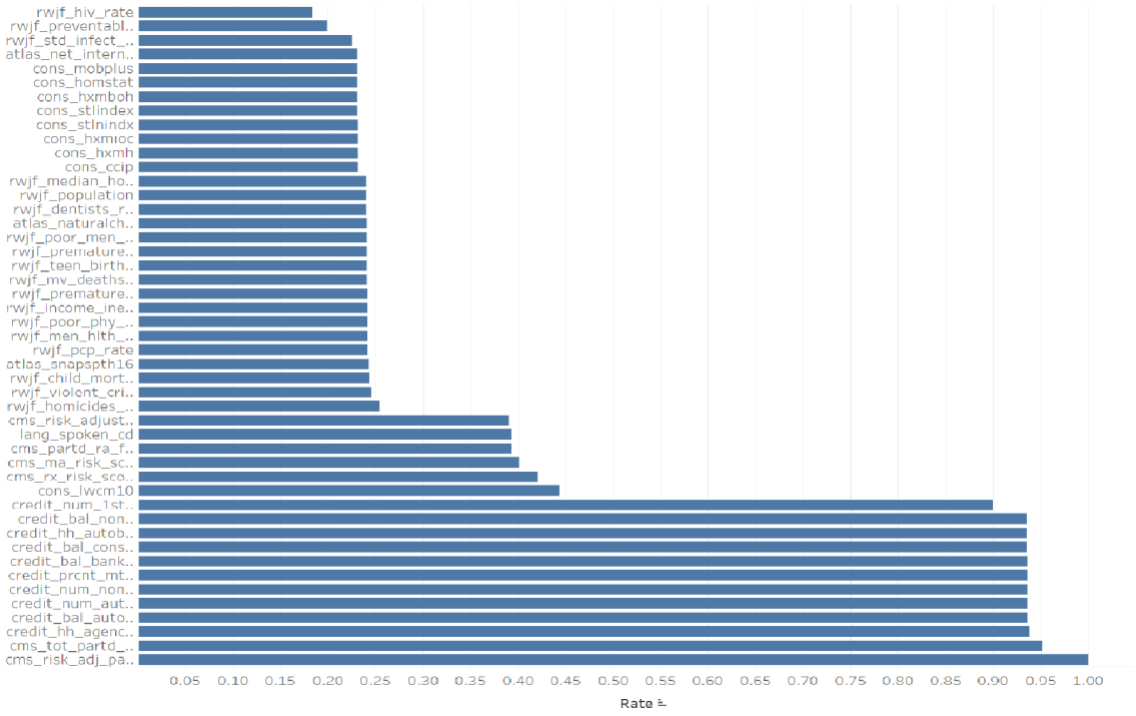


Figure 5. Missing Values

For different types of features, we apply different missing data imputation methods:

Table 2. Missing Data Imputation

| Variables | Description | Replacement |
|---|---|---|
| Categorical variables | For all the categorical and ordinal variables, we kept the null value as a new category *'NA'* to keep the original information. | NA |
| Regional statistics features and credit data | The *atlas, rwjf* and *credit* described the regional statistics for the region where the member was in and personal credit situations, so we imputed the median value for the null values. | Median value |

| | | |
|---|---|---|
| Cost, number of claims and some census data | After observation, the features describing the cost and the number of claims related to a specific disease or clinical diagnosis contain a large number of zeros. We can think that some missing values may also be due to the fact that the member does not have records in this regard, so it was | 0 |
| | collected when the data was collected. resulting in missing values, so we decided to fill all missing data with 0. | |
| Days since last claim for overall claims | We look into the distribution of attributes scaled in days and find that most values are 480 which is maximum values. It's possible because so many members had a long time not doing certain things like going to a pharmacy. | 480 (max value) |

### 3.4 Feature Selection

After data preparation, our columns changed from 881 to 921, the dimension of our dataset increasingly growed. Since using such a huge dataset can cause time costing and overfitting problems, we decided to find key indicators among those 920 features(except response variable *hi_flag*). The most common and effective method to extract important features is using boosted tree models, so we utilized three methods, Gini Index, Random Forest and XGBoost.

We looked into importances greater than zero under the Gini index, the number of variables with zero Gini importance is 522. Random forest selected 684 important features among 920 features. Using XGBoost, we got 499 features with importances greater than zero. After three methods of filtering, we finally got 499 important features based on their intersection and removed 421 unimportant variables. Further work of feature selection was model-specific and conducted later during model tuning.

## 4 MODELING

### 4.1 Model selection

After we selected key factors with importance greater than zero, the model selection was the key to our accuracy of prediction. Our objective is to predict members most likely to have housing insecurity issues under low bias. First, we identified this as a classification prediction question. Second, since we have a high dimensional dataset

(48,300x500), we need to choose a model with good flexibility, high predictive power and easily explainability, so we mainly focused on tree-based boosting algorithms to estimate the probability of housing issues. Therefore, we selected these four models: Random forest, Gradient Boosting Decision Tree, LightGBM and XGBoost to do preliminary prediction and compared their performances after that.

Out of 48000 observations in the original training dataset, we splitted 70% as training data, 30% as testing data. Then, we performed cross- validation with 70% training data and tested three models' performances separately on 30% testing data. The evaluation metric used was the AUC-score, which measures the area under the Receiver Operating Curve (ROC) and generally reflects how well the model can distinguish the classes.

Finally, out of all the models we tested, XGBoost has the best performance of 0.761 in terms of AUC score; Gradient Boosting Decision Tree had an AUC of 0.748; Random forest returned a score of 0.724; LightGBM returned a score of 0.752.

## 4.2 Final model construction

Based on AUC metric in the figure below, the AUC score of the XGBoost Classifier is 0.761 and outperforms the rest, so we decided to use XGBoost to predict on our holdout dataset. Besides excellent prediction performance and fast processing speed, the XGBoost can deal with the imbalanced problem existing in our dataset, where out of 48,300 observations, only 4.4% of members (2,118) have housing insecurity issues, and most members are not suffering from housing insecurity. XGBoost can achieve tuning the training algorithm by "*stratify*" argument to pay more to misclassification of the minority class for datasets with a skewed class distribution. To better analyze the performance of our model, we also calculated the confusion matrix, when we set the threshold to be 0.12, the true positive rate of predications is 90.9%, and false positive rate is 4.71%, which is relatively low. So that the performance of our model is excellent.

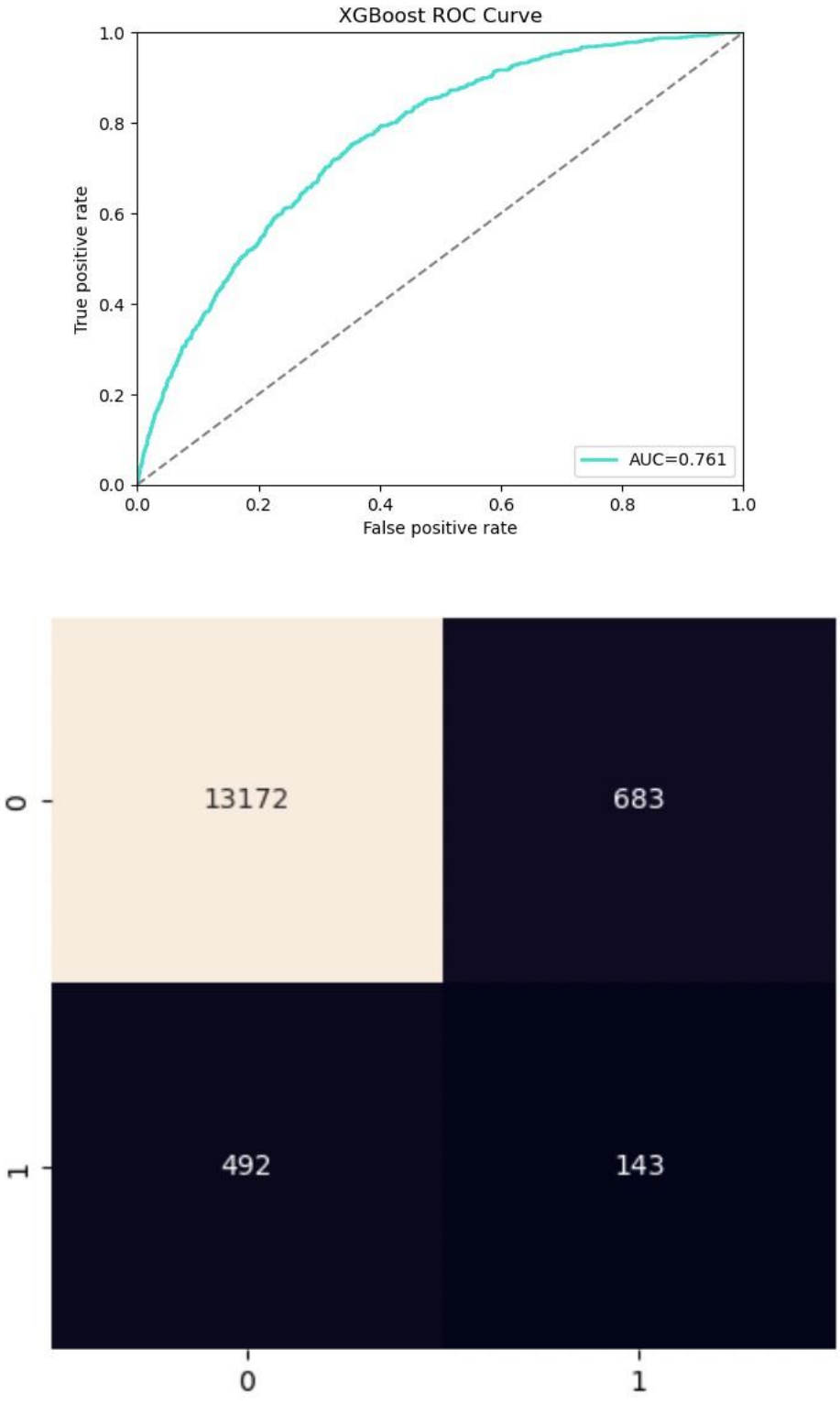


Figure 6. ROC Curve and Confusion Matrix

To improve the model's performance, we did parameter tuning, and the final results are shown below:

1. 'n_estimators': 500. This parameter controls the number of gradient boosted trees, so we set this as a good level to make the model learn.
2. 'objective': 'binary:logistic'. This parameter tells us that our model is binary classification and the outcome is probability.
3. 'learning_rate': 0.05. This parameter is step size shrinkage used to prevent overfitting. After each boosting step, we can directly get the weights of new features, making the boosting process more conservative.
4. 'gamma':0.05. The parameter sets the minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be.[3]
5. 'subsample':0.75. This parameter determines the subsample ratio of the training instances to avoid overfitting.
6. 'colsample_bytree':0.35. This parameter is about the subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed.
7. 'min_child_weight':30. This parameter identifies when to stop tree partitioning. If the tree partitioning results in the sum of weights less than this parameter value, the tree building will give up further partitioning.
8. 'max_depth':6. This parameter is positively related to the complexity of the model, so we set this at a level with good performance but avoiding overfitting.
9. 'seed':1024. Random number seed.

# 5 KEY PERFORMANCE INDICATOR ANALYSIS

## 5.1 Feature Importance

To better understand the model and important features, and drive insights from the model. We looked up the top 30 important features in XGBoost gain importance and top 20 important SHAP values.

- Gain Importance

[3] https://xgboost.readthedocs.io/en/stable/parameter.html

We used the built-in XGBoost feature importance function to get the most important features after tuning and training the model. The calculated numerical value of “gain” to take each feature’s contribution to each tree in the model is the most common method to evaluate the importance of the features in the model. The top 30 important features of gain importance are shown in the following figure.

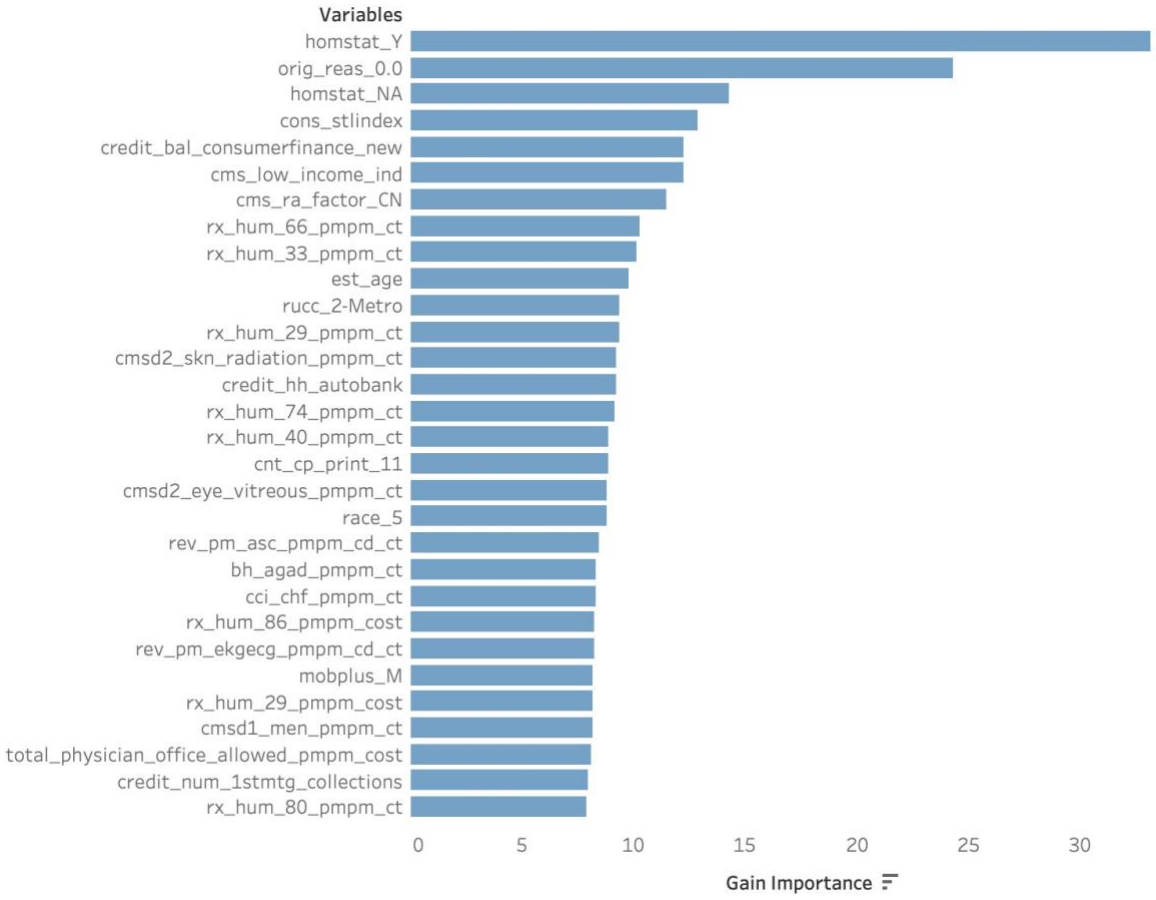


Figure 7. Gain Importance of Features

- SHAP Value

In our feature importance analysis, SHAP is also a well-known method in post-model analysis to compare and analyze the final features, since it generates numeric values which can be used to calculate the important role of the features to the model. The top 20 features of Shap value are shown in the following figure.

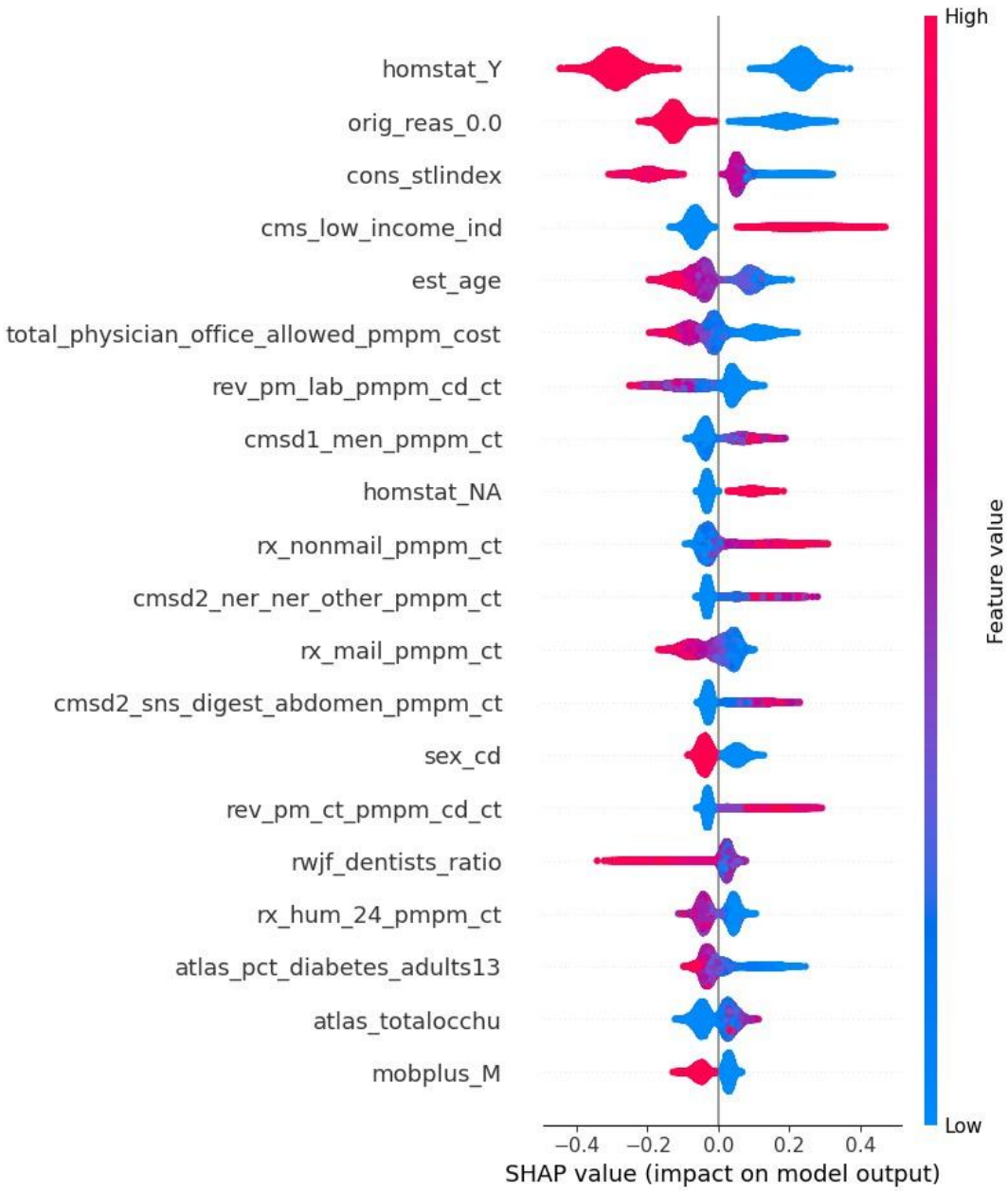


Figure 8. SHAP Value of Features

As can be seen in the gain importance figure and SHAP value figure, some features stand out and have high importance in both figures. And comparing the aspects of all variables, the features can be categorized as the following:

1. The status of the homeowner factor: ‘Homstat_Y’ is the variable ranked first in both figures, which is the most important feature in our XGBoost model. We can see that whether the member is a homeowner is highly related to home insecurity issues, members who are not homeowners are more likely to have home insecurity issues.

2. The original reason for the medicare factor: ‘Orig_rea_0’ variable is the second important feature in terms of both gain importance and shape value. The variable indicating the original reason for entry into Medicare, whether the member entry into medicare is because of old age and not health problems and disability is also important. We can see that if a member purchased the insurance because of their age, it’s less likely for them to have home insecurity issues.

3. Financial-related factors: ‘cms_low_income_ind’(low_income index) and ‘cons_stlindex’(loan index) rank as top5 in both figures, and which other financial-related factors with prefixe of ‘credits’ also shown in top 30 features, which indicate that the economic condition of members also accounts for an importance role in terms of home insecurity issues. Low-income members are more likely living in houses that are not safe. Since cms_low_income_ind variable is a binary variable and easier to identify people with financial issues, we chose this variable to represent financial factors.

4. Health-related factors: We can see that ‘cmsd2’,‘rx’,‘rev’,‘bhi’,‘cci’,‘total’ features also occupy a large proportion, which are originated from member’s Medicare Advantage information, could suggest reasons related to medical or physical conditions. We can see that some health issues are possibly related to the insecurity of the house, including some radiation-related skin issues, eye issues and mental health issues. Among all the health-related factors, we can see that ‘total_physican_office_allowed_pmpm_cost’ indicates the cost that members spent on health issues in the past year, generally it’s related to all other health variables. So we decided to use the ‘total_physican_office_allowed_pmpm_cost’ variable to represent the health factor in the following analysis.

5. Demographic and geographic factors: we can see that ‘Est_age’ is negatively related to home insecurity issues.‘rucc’(Rural Urban Continuum): members living in rural areas are more likely to have home insecurity issues. Additionally, some “rwjf” factors related to physical environment also have a relationship with home insecurity.

## 5.2 Relationship between factors

To further analyze the important features and the relationship between these factors, we first generated the heatmap to see the correlation relationship of each factor. As can be shown in the following heatmap, the original reason of entry medicare is highly positively related to age, and it’s also negatively related to health factors and financial factors;the low income factor is negatively related to home status factor; what’s more, the geographic factor of member has relatively low related relationship with other factors.

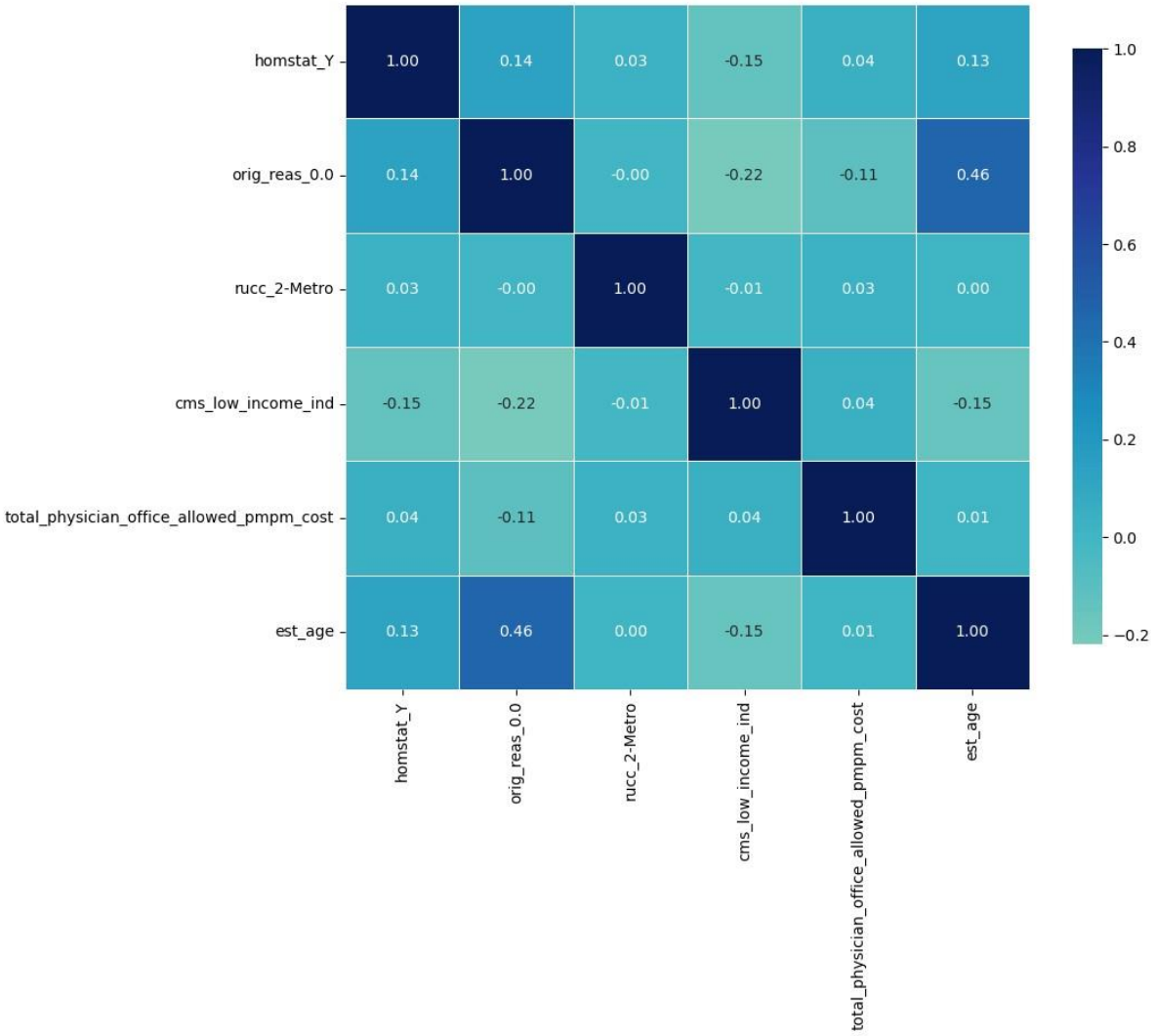


Figure 9. Heatmap of Important Features

Additionally, we used SHAP dependence plots to study the individual effects and interaction effects of key variables.

- Origin_reas_0

The following two dependency plots show the relationship between origin_reas_0 and age, health factors. We can see that the binary variable origin_reas_0 correlates with an increase of age and a decrease of health factors (total_physician_office_allowed_cost). The binary variable origin_reas_0 measures whether a member entry for medicare because of old age, and not because of disability or serious disease. This is aligned with the observation we figured out and information shown in the figure and it is reasonable to explain the negative relationship between age and home insecurity.

Therefore, we can conclude that members purchased insurance because of their age are relatively healthy and older members.

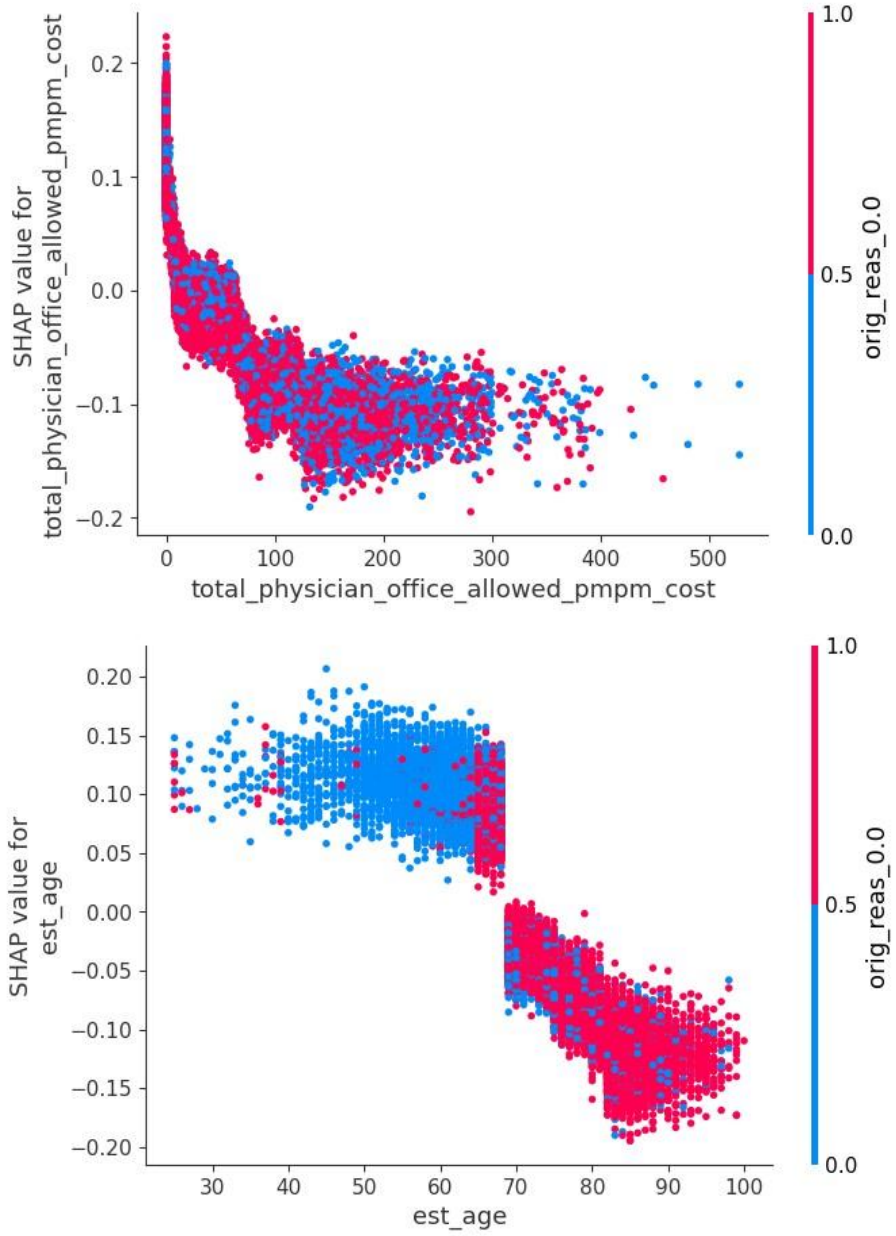


Figure 10. SHAP Dependency Plots

- cms_low_income_ind

Furthermore, we analyzed the relationship between financial-related factors with homestatus factor, health factors. As the dependency plot below shown, the binary variable homestat_Y shows relatively negative relationship with binary variable cms_low_income_ind, more members who have low income not homeowners. But there's still a great group of members who don't have financial issues that are not homeowners. Therefore, financial factor is not the only reason accounts for whether members live in their own house.

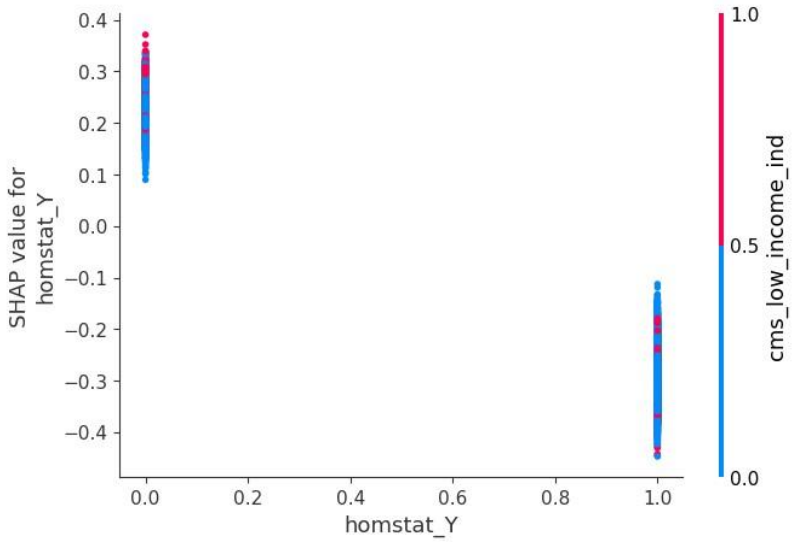


Figure 11. SHAP Dependency Plot

The plot below shows that the binary variable low-income correlates with an increase of the cost in total physician office. It's understandable that low-income people are more likely to have health issues than high-income people.

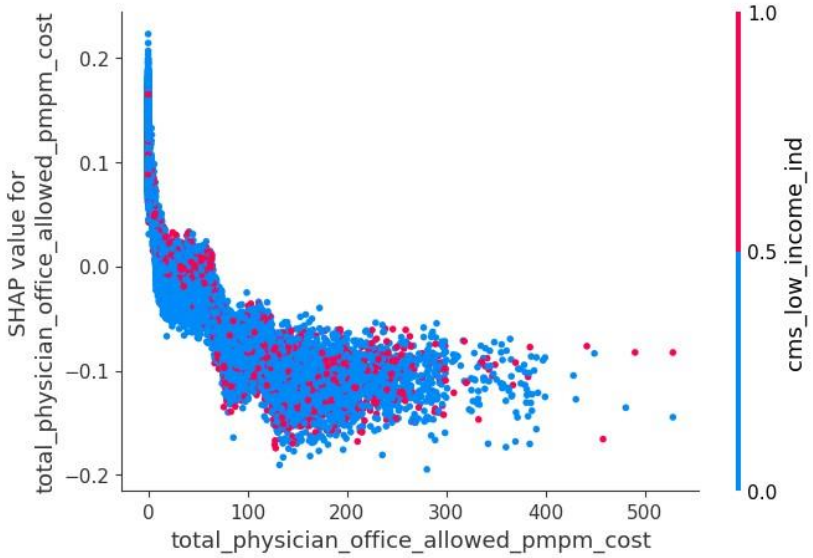


Figure 12. SHAP Dependency Plot

# 6 SEGMENTATION

## 6.1 Segment features

Based on the information research and analysis of features generated from our predictive model, it’s become clear that the home insecurity issue occurs because of different reasons for members. Thus, to dig deeper into the underlying reasons that lead to the house insecurity issues and successfully design appropriate recommendations

for improving house insecurity issues, we need to classify members into different segments.

According to the importance and relationship analysis of features we generated before, we selected the following 3 key binary features in different fields and separated all members into 3 segments:

- homstat_Y: whether the member is homeowner, which indicates the home status
- orig_reas_0.0: whether the original reason for entry into Medicare is Old Age, and not diabled and disease reason, which indicated the health conditions and age of the member
- cms_low_income_ind: Binary indicator that a member is receiving a subsidy from cms, which indicates the economical level of the member

We used the K-means clustering method to segment all 48,300 members in the training data, by evaluating the relative low Inertia score of clustings, we decided to separate all members into 4 segments.

For every segment, we calculated the average of home insecurity flags that indicate the home insecurity level of each segment. At the same time, we also calculated the average value of other important fatcors, including health-related factors and age factors and that might be helpful to understand the reason for the home insecurity problem of different segments.

The summarized conditions and values of each segment are shown in the table below.

Table 3. Clustering Results

| Segments | | | | |
|---|---|---|---|---|
| **Cluster Varibale** | Segement1 | Segement2 | Segement3 | Segement4 |
| homstat_Y | 0 | 0 | 1 | 1 |
| orig_reas_0.0 | 0 | 1 | 0 | 1 |
| cms_low_income_ind | 0.32 | 0.16 | 0.23 | 0.08 |
| **Group size** | | | | |
| number | 8,228 | 13,329 | 6,713 | 20,030 |
| percentage | 17.0% | 27.6% | 13.9% | 41.5% |
| **Other factors** | | | | |
| Age | 65.56 | 74.08 | 66.96 | 75.08 |
| Health Factor | 60.92 | 47.57 | 68.61 | 53.33 |
| **Home Insecurity** | | | | |
| hi-flag | 9.94% | 4.91% | 4.48% | 1.72% |

The figure above describes the result for each segment, and each group is classified as homestatus, original reason and low-income variable, the average of home insecurity flag is different from each segment and they are also different in age and health factor. Therefore, the segmentation of all members is reasonable and meaningful to help better understand the reason that would be related to the home insecurity problem of members.

## 6.2 Segments Analysis

Segment 1:  Non-homeowner and Low income members

This segment represents those who are not homeowners of their living space and have the greatest low-income issues, which accounts for 17% of all members. There are 9.94% of them who have home insecurity issues, which is the highest compared to other segments, indicating that those low income and non-homeowner members are most likely to suffer from housing insecurity problems.

To understand why the segment is most likely to suffer from home insecurity, we do some investigation of secondary research.

According to the survey by the US census Bureau, the number of families facing severe housing burdens, meaning they spend more than 50% of their income on housing, has increased over the last decade. Especially in 2020, shown in the figure below, there

were about 23% renters in the U.S. spending greater than 50% of their income on housing costs. U.S. housing prices shot up 18.8% in 2021, the highest calendar year increase in 34 years of data[4]. Residential rents typically follow home prices, this January's average rents were 15.2% higher than last January's. Rents skyrocketed in 48 of the 50 biggest U.S. metro areas.

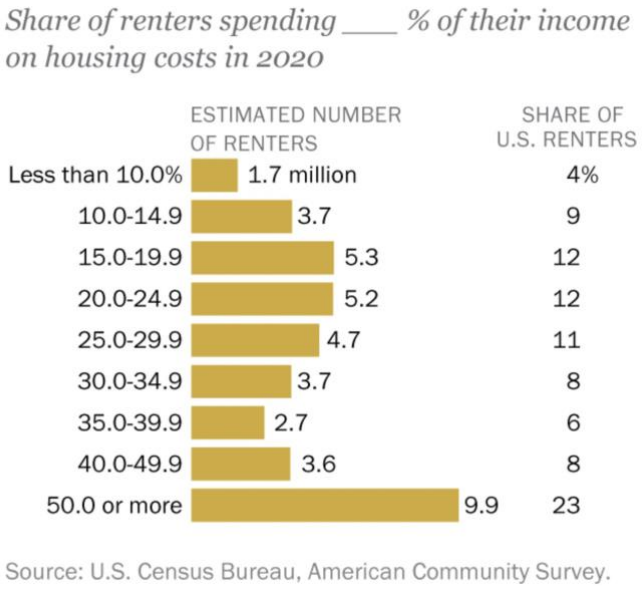


Figure 13. Share of renters spending their income on housing in 2020

Therefore, we can conclude that for this segment, low-income is the main reason that forces them to live in an insecure place. The main pain point of this group of people is the increasingly common issue of low income level and the increase of housing prices and renting cost.

Segment 2: Non-homeowner and elderly members

This segment represents those who are not homeowners of their living space but don't have low-income issues, which accounts for 27.6% of all members. 4.91% of them have home insecurity issues, which is the second highest among all segments. which is relatively high and close to the second segment. Additionally, the average age of this group is 74, and their health factor is the lowest among all segments, indicating that they are a group of healthy elder members.

To understand why this segment also has a high probability of facing home insecurity problems, we searched for more information. Indicated by the State of the Nation's Housing report, adults aged 55 and over contributed about two-thirds of rental housing

[4] https://www.pewresearch.org/fact-tank/2022/03/23/key-facts-about-housing-affordability-in-the-u-s/

growth from 2004-2019. This age group now constitutes 30 percent of all renter households, and over 13.2 million households. Since the age distribution of Humana's members is concentrated in 60 to 80 years old, most of them are retirees. Therefore, the reason why this part of non-homeowners who have no income troubles choose not to buy a house is that as a retiree living alone, renting can reduce the cost and energy of house maintenance and repair. Furthermore, they tend to move closer to their families as they are getting older, where their children work and live, mostly in cities. Therefore, high housing and population densities in cities, coupled with noise, air and water pollution, are the main reasons attributed to their housing insecurity.

Segment 3: Homeowner and Low income unhealthy members

This segment represents those who are homeowners of their house but suffer from low-income issues, which accounts for 13.9% of all members. There are 4.48% of them having home insecurity issues, which is relatively lower than the second segment. The health factor is the highest among all segments, indicating that they are a group of unhealthy poor people.

As far as we are concerned, for these low-income homeowners, their low income drives the low quality of their living conditions and also worse health conditions. The quality of housing drops because low income people are very likely to live in a violent, messy and unstable neighborhood. Their housing security and life standards cannot be ensured at a high level. On the other hand, the lack of money to repair their own houses or apartments would negatively affect their living conditions.

Segment 4:Senior high income and homeowner members

The segment 4 represents those who are homeowners and having no income issues. This segment accounts for 41.5% of all members, which is the biggest segment. Among those high income homeowners, only 1.72% of them have home insecurity issues, which is the lowest. The age of this group is 75 so that they are the oldest members.

To understand why these groups of segments may encounter issues of home insecurity, we calculated some important healthy variables and found that some specific health variables stand out. As can be seen in the following table, segment 4 are more likely to have radiation-related skin health issues, eye and adnexa issues and some mental health issues that come from the insecurity of their home.

Table 4. Relationships of Segments and Specific Diseases

| Variables | Feature Description | Segment 1 | Segment 2 | Segment 3 | Segment 4 |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| cmsd2_skn_radiation_pmpm_ct | diseases of the radiation-related skin claims | 0.013 | 0.006 | 0.009 | 0.020 |
| cmsd2_eye_vitreous_pmpm_ct | diseases of the eye and adnexa claims | 0.011 | 0.006 | 0.007 | 0.014 |
| rx_hum_66_pmpm_ct | mental health related drugs claims | 0.005 | 0.005 | 0.006 | 0.006 |

## 7 RECOMMENDATIONS

Based on our segmentation, we have four main clusters of members based on economic factors and home status. Since members experiencing housing insecurity are not homogeneous and may have different drivers for their insecurities, we need to specify their reasons and then give solutions to each segment. To improve housing quality for most people, we ranked our clusters of members using urgency and feasibility metrics, and then we gave specific solutions to target segments, covering as many members from diverse segments as possible.

Our priority is to deal with housing insecurity faced by non-homeowners. Since they are most likely to have housing insecurity issues, which is far beyond possibilities of housing insecurity issues from other groups. Second, we focus on financial deficiency, since financial problems are the second major factor for housing insecurity based on our analysis. After that, we want to consider health-related issues, since people in one of our segments have some common issues, especially skin, ophthalmology issues and mental health issues. They have housing instability and other issues, neither because of low income nor because of not owning a house. They are likely to be affected by potential housing issues especially for people who have some kinds of disease histories related to skin and ophthalmology issues. Therefore, in the long term, we need to give personal care to those people, and improve their housing quality to avoid housing insecurity.

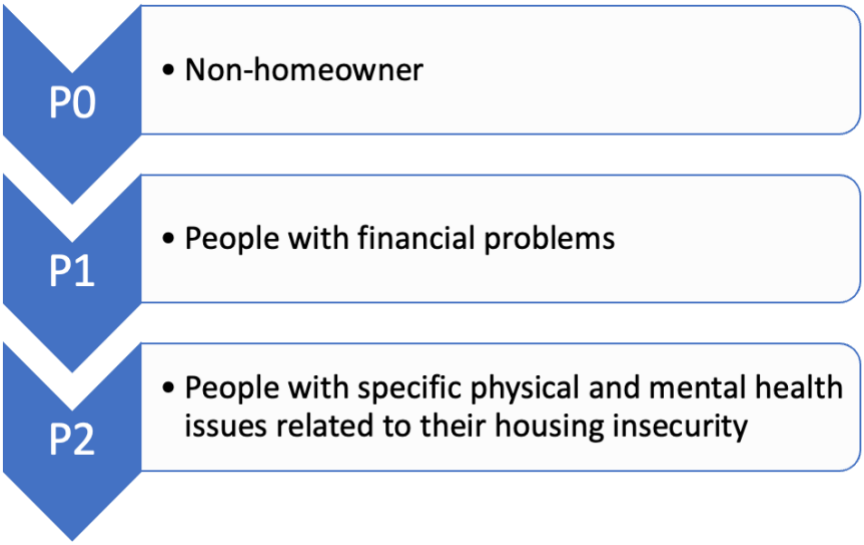


Figure 14. Prioritized Strategy Roadmap

## 7.1 Strategy program

### 7.1.1 Segment 1:  Providing financial assistance

As we analyzed above, the reason for this segment that is most likely to suffer from housing insecurity problems is low income. The number of families facing severe housing burdens, meaning they spend more than 50% of their income on housing[5]. Our mission is to address housing affordability, which is the most cost-effective way of lifting people out of poor living conditions, to enhance people's living standards.

For low-income members, we suggest providing financial assistance for them and helping them with getting a new living place, on the other hand, we offer them professional help to find a temporary living place while they can get back on their feet.

Therefore, we give our recommendations: (1) Humana could collaborate with banks and financial institutions to provide a housing subsidy program, low property tax programs and home loan programs. (2) Humana collaborates with housing agencies to facilitate opportunities and access to housing resources and assist members with housing search and application progress, offering easier access to economical and affordable housing.

Implementation:

1. Help members to connect with housing finance agencies: with the referral of Humana, members under financial burdens can gain financial assistance from federal agencies since housing finance agencies can make low-rate housing loans through the sale of taxable and tax exempt bonds.
2. "Good Home Selection Plan": This plan is to give an economical plan to move to a new and secure place. After members get incentives from agencies, Humana cooperates with real estate agents to introduce high-quality but cost-efficient housing to them, which helps save time and financial cost to look for a new living site.

[5] S&P CORELOGIC CASE-SHILLER INDEX REPORTS 18.8% ANNUAL HOME PRICE GAIN FOR CALENDAR 2021

### 7.1.2 Segment 2: Providing convenient home maintenance and repair

As we have analyzed before, high housing and population densities in cities, coupled with noise, air and water pollution are the main reasons attributed to their housing insecurity.

Improving housing insecurity for this group of people focuses on addressing their concerns and worries related to the inconvenience of housing repair and maintenance. Therefore, we suggest that Humana can provide members with easy home problem reporting and repairing services by collaborating with home maintenance and repair contractors.

Implementation:

1. Add a “one click repair” function to the website or app. When the house needs to be maintained or repaired, members can choose the specific place and facilities that need to be repaired.
2. After receiving the member's repair request, the collaborator will allocate the corresponding maintenance personnel to deliver on-site services.

### 7.1.3 Segment 3: Improving living environment and housing quality

Based on segmentation analysis before, the quality of living environment and housing conditions are the main factors that lead to the home insecurity of this segment. To address low housing quality problems for them, we identified the main factor is financial deficiency. Thus, our recommendations are from two aspects: (1) to improve their environments, (2) to improve their housing quality. The first aspect is restrained from local federal financial abilities, so we focus our recommendation on the second aspect that Humana cooperate with repair companies and give fixing solutions at a good price.

Implementation:

1. Screening: Humana regularly checks low-income homeowners’ housing quality, from external facilities and internal furnitures.
2. Help maintain members’ homes: Humana incorporates with housing repair companies to give subsidized housing repair assistance for members whose houses need urgently fixing and decorating.

### 7.1.4 Segment 4: Establishing health management system, providing housetesting and disinfecting

According to our model results, combined with the importance of features, various health issues of Humana members are strongly related to housing insecurity. People experiencing physical problems like eye, skin diseases, or mental health issues are

likely to have housing insecurity issues. Nearly 80 million Americans have medical debt, which complicates taking care of themselves[6]. All the stress of medical expenses, lack of payment options for services, and housing insecurity can contribute to chronic diseases and have a serious impact on overall physical and mental health. A 2015 study reported that housing insecurity is associated with physical and mental health problems, and that groups with housing insecurity are more likely to avoid medical care and have health-risk behaviors and outcomes. After adjusting for demographic and socioeconomic indicators, respondents with housing insecurity were more likely than those without housing security to report: delayed doctor visits, poor or good health, healthy within the past 30 days Poor condition and poor mental health for 14 days or more, limiting daily activities[7].

- Physical health issues: We observed that members with housing insecurity had higher rates and costs of claims related to eye diseases, skin diseases and subcutaneous tissue surgery. On the one hand, members with these physical health problems are more likely to be identified as having potential housing insecurity problems, such as the housing environment, poor sanitation, and high density of people, ophthalmology, dermatology caused by poor living environment , chronic diseases related to the respiratory system and allergic diseases; on the other hand, housing insecurity will further affect or even aggravate the health problems of these groups, resulting in a vicious circle.
- Mental health issues: Housing insecurity is associated with adverse mental health effects. On the one hand, groups affected by housing insecurity have a higher risk of depression, anxiety, and even suicide. These groups are 6-10 times more likely to have poor mental health than the general population. On the other hand, it is more difficult for people with poor mental health to actively deal with housing insecurity, thus further exacerbating the negative impact of housing insecurity on their mental state.

Implementation:

1. Establish a complete health management system for members, encourage members to carry out daily health management under the guidance of Humana, and report the health problems they face. The system can encourage members to regularly record and actively manage their own health. For example, if a

---

[6] Stahre M, VanEenwyk J,Siegel P, Njai R. Housing Insecurity and the Association With Health Outcomes and Unhealthy Behaviors, Washington State, 2011. Prev Chronic Dis 2015;12:140511. DOI: http://dx.doi.org/10.5888/pcd12.140511.

[7] Juli Carrere, Hugo Vásquez-Vera, Alba Pérez-Luna, Ana M. Novoa, Carme Borrell. Housing Insecurity and Mental Health: the Effect of Housing Tenure and the Coexistence of Life Insecurities. J Urban Health (2022) 99:268-276. DOI: https://doi.org/10.1007/s11524-022-00619-5

member has a chronic disease, he or she should regularly register some relevant health data in the system, and then conduct daily self-management according to some medical guidance from Humana, such as regular inspections at designated institutions, taking medicines according to courses of treatment, and daily exercise, and uploading the records to the system, so that Humana can manage members' health more scientifically and efficiently. Insist on uploading health data records and self-management data, you can get some incentives, such as reducing insurance premiums.

2. For physical health problems caused by housing insecurity, such as skin diseases, respiratory diseases and chronic allergies, Humana can regularly provide its members with some tests for harmful substances in the house (such as formaldehyde), house cleaning, and insects and mites removal services.

## 7.2 Cost & Effectiveness Analysis

After we segmented members, we need to quantify the cost of our recommendations, and if the total cost of implementation for each recommendation can offset or be smaller than total insurance claim fee, these recommendations can be cost-efficient.

The total insurance claim is $319.36 on average for one member per month. For each year, Humana has to pay for $319.36 * 48300 * 12 = $**185,000,000.** We assume that the average combined loss ratio is 40%. Therefore, the approximate claim is $**74,000,000** a year for 48,300 members.

For our recommendations, the calculation of total cost should be:
Total cost = N * [ S1 * C1 + S2 * C2 + S3 * C3 + S4 * C4 + Ce ]

Where:

(1) N = total number of members
(2) Si = percentage of members who are in segment i. (i=1,2,3,4)
(3) C1 = cost per member who needs financial assistance and new site searching.
(4) C2 = cost per member who gets professional housing consulting.
(5) C3 = cost per member who requires house repair and maintenance.
(6) C4 = cost per member who needs regular health tests.
(7) Ce = other cost, like labor cost for implementations.

C1: Helping members to contact housing financing institutions and the "Good Home Selection Plan" launched in cooperation with housing agencies will be public welfare

projects. Therefore, we mainly include the labor cost of investment, and this part will be put into Ce for estimation.

C2: The input cost of this part is mainly divided into: the investment in upgrading the functions of the website and the app, and the cost of cooperating with a house repair company to provide members with affordable repair services. Last year, the average household spent $3,018 on maintenance costs and $2,321 on emergency repairs, according to Angi's State of Home Spending Report[8]. According to our segment analysis, we will mainly offer 48300 * 27.6% * 4.91% equal to 655 members (328 families) using maintenance and repairs approximately about $**1,751,192.**

C3: For segment 3 people, we provide a 30% subsidy for maintenance and repair of their housing: Therefore, we need to provide 48300 * 13.9% * 4.48% equal to 301 members (about 150 households) to provide about **$240,255** in subsidy

C4: The cost of the main proposed measures for the fourth segmented population is concentrated in two aspects:

- Cost of building a membership management system and labor costs,on the basis of combining existing medical resources and databases: Due to the company's existing medical resources and database construction needs to be further evaluated, we are temporarily unable to evaluate the total cost of the project.
- Cost of house cleaning and improvement services: Industry average level shows that some cleaning companies might charge by the square foot, but they normally will just want to look at the house and determine how hard it will be to clean. The average price for a standard house cleaning is from $0.10 per square foot to $0.17 per square foot. This means for a 2000 square foot house you can expect to pay between $200 to $340[9]. According to our segment analysis, On average, there are 41.5% * 1.72% of Humana members, that is, 48300 * 41.5% * 1.72%, which equals 345 people（about 173 families） facing the need for housing improvement. Calculated according to the lowest average price in the industry, the average 100 square foot per house and the average frequency of once every six months, Members cost about $34,600 per year in housing improvements. Humana will cover 30% of the subsidy to serve these members, so the cost will approximately be **$10,380**

[8] https://www.sofi.com/learn/content/most-common-home-repair-costs/

[9] https://fastmaidservice.com/house-cleaning-prices/

Ce: The labor cost related to one member, since the number of members is larger than that of employees in Humama, so the labor cost divided by number of members is, we assume, approximately zero, compared to other costs.

Therefore, the least total cost of recommendations is $**2,001,827**, smaller than the assumed claims a year for Humana. Therefore, the recommendations are cost effective.

## 8 CONCLUSIONS

In predicting members who are most likely to be experiencing housing insecurity issues, we first selected the most important features through Gini Index, random forest and XGBoost. Then we applied Random Forest, Gradient Boosting Decision Tree, LightGBM and XGBoost to do preliminary prediction and compared their performances and corresponding AUC. According to our model, XGBoost has the best performance with an AUC of 0.761, outperforming the other models. In the segmentation analysis, based on their home status and income level, we used the K-means clustering method to segment all 48,300 members into four groups, which covered all the members struggling with housing insecurity with low overlap among each other. For each of the four groups, we put forward targeted and personalized recommendations, including providing financial assistance, offering convenient home maintenance and repair, improving living environment and housing quality, establishing health management system and providing house testing and disinfecting, with these measures Humana can improve housing quality and seek the greatest benefit for its members.