Table of Contents

I. Executive Summary

II. Case Background

As one of the top five health insurance companies in the U.S., Humana believes in a whole person health approach. As opposed to providing disconnected clinical needs, whole-person health involves supporting individuals by addressing their physical, mental, and social health in a comprehensive way. Health and disease are interconnected. A myriad of factors, including an individual's genetic makeup, life choices, and socio-economic aspects, contribute to one's long-term health. By incorporating a holistic approach to care for its members, Humana provides a system that improves the overall health outcomes of over 10 million people in the United States.[1]

One of the important social factors that impact health is housing insecurity. The U.S. Department of Health and Human Services defines housing insecurity[2] as the lack of security due to high housing costs relative to income, poor housing quality, unstable neighborhoods, overcrowding, or homelessness. In 2019 before the COVID pandemic, nearly 37 million households, including renters and homeowners, were cost burden – spending more than 30 percent of their incomes on housing.[3] In January 2020, there were 17 per 10,000 people experiencing homelessness in the U.S.[4] The COVID pandemic further exacerbated housing insecurity issues across the country due to soaring housing prices and loss of employment. Certain vulnerable groups, including racial minorities, young children, and older adults, have been disproportionately impacted by housing instability.[5]

---

[1] According to its 2021 annual report, Humana serves over 17 million medical members. Its medical membership includes nearly 8.6 million are Medicare members, and 4.9 million are enrolled in a Medicare Advantage plan. See Humana, "Humana Expanding Medicare Advantage Health Plans in 2022 to Address Beneficiaries' Most Important Needs, Delivering Predictable, Affordable and Understandable Health Care," October 1, 2021, available at https://press.humana.com/news/news-details/2021/Humana-Expanding-Medicare-Advantage-Health-Plans-in-2022-to-Address-Beneficiaries-Most-Important-Needs-Delivering-Predictable-Affordable-and-Understandable-Health-Care/default.aspx#gsc.tab=0; Humana, 2021 Annual Report, available at https://www.annualreports.com/HostedData/AnnualReports/PDF/NYSE_HUM_2021.pdf.

[2] The U.S. Department of Health and Human Services uses the term, housing instability. See Johnson, Amy and Alicia Meckstroth, *ASPE*, "Ancillary Services to Support Welfare to Work," June 21, 1998, available at https://aspe.hhs.gov/reports/ancillary-services-support-welfare-work.

[3] Office of Disease Prevention and Health Promotion, *U.S. Department of Health and Human Services*, Healthy People 2030 Housing Instability, available at https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries/housing-instability.

[4] National Alliance to End Homelessness, "State of Homelessness: 2022 Edition," 2021, available at https://endhomelessness.org/homelessness-in-america/homelessness-statistics/state-of-homelessness/.

[5] Parker, Jim, *Hospice News*, "Housing Insecurity a Threat to Seniors, Poses a Risk for Hospice Providers," July 7, 2022, available at https://hospicenews.com/2022/07/07/housing-insecurity-a-threat-to-seniors-poses-a-risk-for-hospice-providers/; Pagaduan, Julie, *National Alliance to End Homelessness*, "Millions of Americans Are Housing Insecure: Rent Relief and Eviction Assistance Continue to Be Critical," November 9, 2021, available at https://endhomelessness.org/resource/housing-insecurity-rent-relief-eviction-assistance/#citation2.

Housing insecurity has a direct impact on an individual's health care choices and outcomes. Living in a substandard place, such as sharing a space with too many people or having inadequate utilities, is likely to cause chronic stress which negatively affects one's mental well-being. Poor environmental conditions, such as exposure to lead, dust, or toxic chemical, lead to higher risk of suffering hypertension, heart damage, respiratory issues, and neurological impairment.[6] The lack of affordable housing often contributes to financial strain taking away resources one would put towards necessary medical care. It is crucial to address the challenges one experiences in housing earlier on to prevent severe and chronic illnesses.

Humana focuses on five social determinants of health, including housing insecurity, to improve the healthcare needs of its members. In its member screening from November 2020 to January 2021, Humana found that "[f]or Humana MA members, 21% report having one or more housing quality issues, which could include pests, mold, water leaks and other issues."[7] In other words, the number of members experiencing housing insecurity issues is substantial. With a better understanding on the housing challenges its members face, Humana can connect each of its members to the needed clinical or non-clinical care through its extensive healthcare networks and community partnerships. More importantly, it can help its members prevent or address health issues at an early stage and improve its forecast on medical costs at the individual level.

---

[6] Humana, "Improve outcomes through wholeperson health: A guide to addressing social determinants of health in patients," 2021, available at https://populationhealth.humana.com/wp-content/uploads/2021/12/Physician-Guide-to-Address-SDOH.pdf.

[7] Humana, "Improve outcomes through wholeperson health: A guide to addressing social determinants of health in patients," 2021, available at https://populationhealth.humana.com/wp-content/uploads/2021/12/Physician-Guide-to-Address-SDOH.pdf.

III. Business Problem

The objective of the case competition has two parts: (1) to identify Medicare members who are most likely to be struggling with housing insecurity issues using a predictive model; and (2) to propose business solutions that would help these members achieve their best health outcomes.

The performance of the predictive model is evaluated by the ROC (receiver operating characteristic) curve and the AUC (area under the ROC curve) metric. The ROC curve describes the true positive rate (i.e., correctly predicting housing insecure individuals) and false positive rate (i.e., predicting housing insecure individuals to have no housing issues) at different classification thresholds for the model, and the AUC metric dictates the accuracy of the model.

The result of the predictive model should reflect fairness and mitigate bias inherent in the data. The ability to maintain fairness in the model is evaluated by the disparity score, which assesses whether our model takes the assumption of equal opportunity into account. Equal opportunity assumption requires that positive outcomes are independent of each sex-race combination. In other words, given a group of people who struggles with housing insecurity issues, regardless of race and/or sex, the probability of any individual in that group being predicted positively should be similar.

IV. Descriptive Analysis

1) Data Overview

The training data describes 48,300 Humana MAPD members with 881 characteristics, and the testing data contains 12,220 members with 880 characteristics. While the training data has the target variable – whether a member experiences housing insecurity in the past year, the 880 characteristics are the same across the two datasets.

Data is heavily skewed towards members who do not have housing insecurity issues. There are approximately 96% of members who do not experience housing insecurity. The imbalance nature of the data is accounted for in our model, which will be discussed in later section, to avoid bias in over-prediction.

The rest of 880 features include, but not limit to, demographic information, medical and prescription claims, credit and consumer data, and regional-level data on social determinants of health. The majority volume, namely 64%, of medical claim information consists of claim counts based on two standard categories, CMS diagnosis codes and a Charlson Comorbidity Index. A detailed list on initial categorization of the features is as follows:

1. Demographic information, including data from CMS (21 features)

2. Credit and consumer data (19)

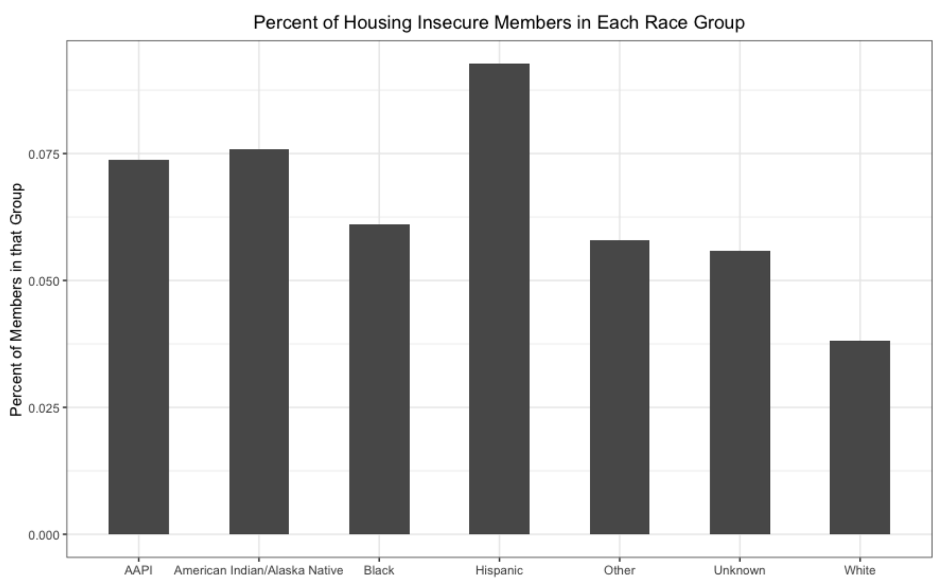
3. Medical claim counts and costs, including behavioral health conditions (511)

4. Pharmacy claim counts and costs (234)

5. Robert Wood Johnson Foundation and Rural Atlas SDOH data (30)

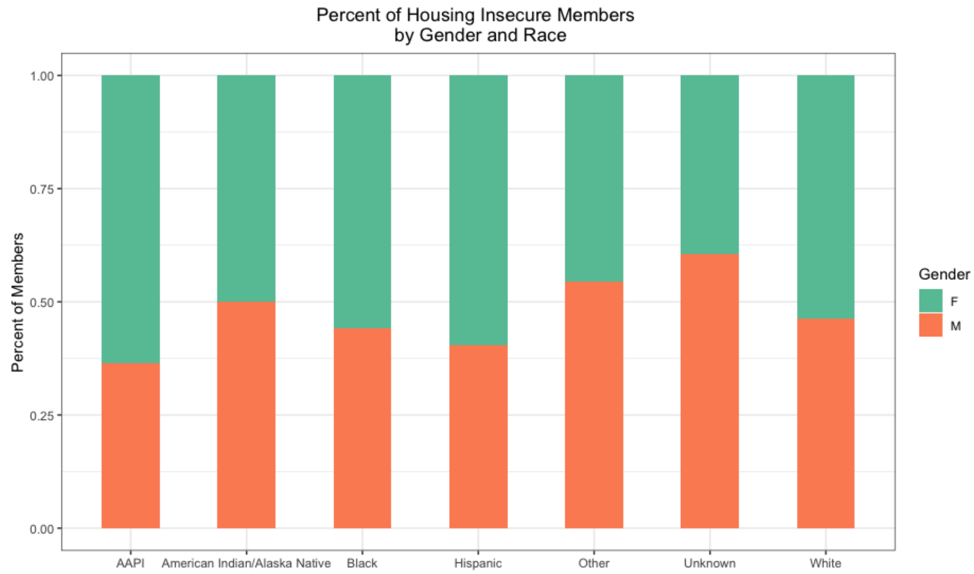
6. Humana outreach features (65)

In the training data, there are 2,118 members classified as housing insecure, and members are categorized into seven races[8]. White population makes up nearly 80 percent, following by Black population at approximately 16 percent. Each of the other five racial categories makes up less than 2 percent. Across each racial group, the proportion of housing insecure and housing secure members varies. Compared to other groups, the Hispanic group has the highest portion of housing insecure members at 9.3 percent. The AAPI and American Indian/Alaska Native groups have approximately 7.5 percent, whereas the white group has the lowest portion of housing insecure members at 3.8 percent.

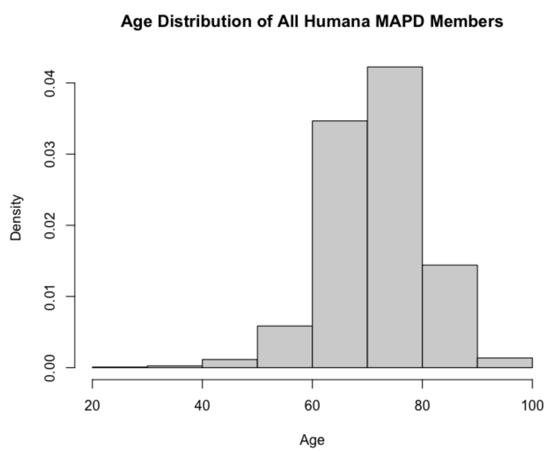


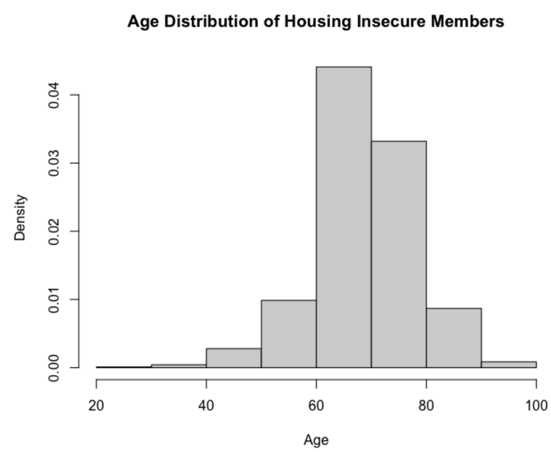
The gender distribution of housing insecure members within each racial group also varies. While the white housing insecure members has a relatively even distribution between female and male, the Hispanic, Black, and AAPI groups have a notably higher proportion of female members who struggle with housing instability, at 60 percent, 56 percent, and 64 percent respectively. However, the sample size of the AAPI group who struggle with housing issues is 22 members, which is low as compared to the sample sizes of the white, Black, and Hispanic groups.

[8] There are 7 members with "*" as the value under the CMS_RACE_CD field. These are recoded as "Unknown" to keep the data on these members. This step maintains the general racial distribution of the data since these 7 data anomalies and the "Unknown" racial category are substantially small relative to the total number of members.



Due to the nature of the age requirement for Medicare plans, more than 85 percent of the members are seniors with the age of 60 and older. While members with the age of 70 to 80 make up the highest percentage of all members, members with the age of 60 to 70 make up most of the housing insecure sample.

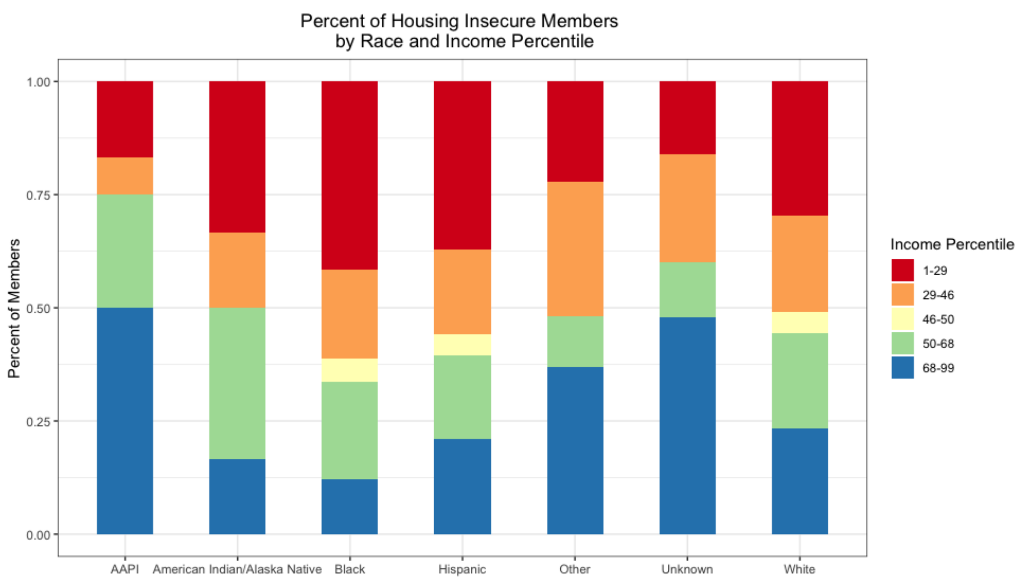




The distribution of housing insecure members with disability or no disability are comparable among racial groups with a reasonable sample size. Within the Hispanic (99), Black (470), and White (1,429) housing insecure members, the proportion with disability ranges from 53 percent to 57 percent. The housing insecure members with disability has an approximately even distribution among gender, with

51 percent of female and 49 percent of male. In addition, the majority of the housing insecure members with disability is from the age of 25 to 66 at approximately 62 percent. Out of housing insecure seniors of the age of 67 or older, the members who also have disability consist of at least 20 percent, which is a considerate portion.

Housing insecurity issues can be closely tied to financial instability. In the training data, 23 percent of all members are associated with missing income percentile information. For the 1,293 out of total 2,118 housing insecure members with known income percentiles, their income distribution can be further broken down by racial groups to detect visual differences in financial situations among different racial groups. The bar graph below illustrates that compared to their white counterparts, a higher proportion of black housing insecure members is in the lowest income percentile from 1 to 29 percent. The proportion of lower income percentiles among Hispanic housing insecure members also follows a similar trend as black housing insecure members. The sample sizes for groups that are other than white, black, and Hispanic (under 30 members) are too small to draw conclusive insights.



2) Data Preparation

Out of the 880 individual member characteristics that are predictors for housing insecurity, 47 characteristics, or approximately 5 percent, have over 15 percent of values that are missing. Most of the characteristics that have between 23 and 45 percent of missing values describes individual consumer data, such as homeowner status and income percentile, and county-level Robert Wood Johnson Foundation SDOH. The characteristics with over 90 percent of missing values are mainly individual credit information, such as the number of severe derogatory accounts and the number of loans that are past due. Based on further investigation, these features with a high proportion of missing values are kept for modeling because of the important information they contain.

To retain as many useful features as possible, data are extrapolated to fill in the missing values. For categorical variables, such as an individual's homeowner status or spoken language, missing values are replaced to be the "Other" or "Unknown" category, which has already been an existing category for those features. In addition, for numeric variables, missing values are replaced with the median of those variables. Such data extrapolation may mask associations among certain strata in the real-world population. However, using the median of the numeric features of the entire training sample without injecting assumptions on certain demographic groups may be the most reasonable approach to mitigate data biases for modeling.

Furthermore, categorical variables are transformed into sets of indicator variables for the model. This data transformation step increases the total 881 features to 918 features, which are fed into the modeling step.

Features in the data inherently have different units of measure. For example, the medical claim costs are in dollars while medical claim counts are in ones. Other features, such as income percentile and migration rate, are in percentages, whereas some, such as diabetes index and health management index, are unitless. A standardization of feature scales is independently employed to normalize each feature scale with a mean of 0 and a standard deviation of 1. This ensures that the features, which are measured at different scales, contribute equally to the model fitting process.

V. Predictive Modeling

1. Model selection

After studying, cleaning and preparing the data, we now can use the data for training. In order to identify members most likely to be experiencing Housing Insecurity issues, we can first use a classifier model to classify the observations, then from the model we predict the probability of each observation. Finally, we rank these probabilities in descending order to show that the higher the probability, the more likely that member will experience Housing Insecurity.

Traditional statistical models like Logistic regression and Lasso regression may not be our best option due to the classification nature of the target variable (housing insecurity flag), the vast amount of data (around 48300 observations with 881 features each), and the non-linear high dimensional interactive relationship between the features and the target. Tree-based models would be a better fit for our classification model because of their tendency to group characteristics and samples together and their great tolerance for non-linear correlations.

Decision tree is the most straightforward model among tree-based models, although it may perform poorly due to bias, small sample size, and high volatility. On the basis of bootstrapped training data, bagged trees construct several decision trees simultaneously, potentially enhancing model performance. By utilizing a method to de-correlate the trees, random forest, in contrast, outperforms bagged trees. In random forest models, split candidates are selected at random from the whole collection of predictors each time a split in a tree is considered. As a result, when we average the trees, random forest models can de-correlate the trees and generate a reduced variance. Boosted trees function similarly to bagged trees, with the exception that each tree is developed in a specific sequence, drawing on knowledge from earlier trees thus correcting each other's errors. This is useful in our instances since gradient boosting trees could be more effective and accurate than random forest models, as a gradient boosting classifier is capable of capturing complex patterns in the data.

After testing all the above-mentioned models, we received an empirical experiment result that all the models have very high accuracy score. However, when we look at the confusion matrix, almost all models predict none to very few members with flag = 1 while most of the predictions are with flag = 0, meaning that our models do not predict correctly even though the accuracy score is high. One possible reason for this problem is that the dataset is highly skewed to begin with: the number of flag = 1 is only 4% of the total observations (2118 out of 48300), making classification task more complicated as any model that predict decently a high number of flag = 0 will have a high accuracy score.

2. Resample

In order to work with this highly imbalanced dataset, we are using **resampling** as a technique to improve our HI to non-HI ratio. Resampling is the process of drawing repeated samples from the original dataset. The intuition behind resampling methods is that it creates "similar" cases for our data classes in order to render the data representative of the population we wish to investigate, and therefore feed the algorithm enough data to output more accurate results (when the data is not enough). From the given dataset, we take random draws of the non-dominating class and create "fake" copies to match the number of cases in the dominating class. In this situation, we are essentially making copies of the data and using those to train our model.

On applying the resampling technique on the dataset and using the outcome to train our models, the results are in line with our previous intuition: Gradient Boosting Classifier gives better prediction results than both Random Forest and Decision Tree methods.

3. Parameter tuning

With all factors considered and numerous training session results, we decided to use the Gradient Boosting Classifier (GBM) package because of its superior performance and relatively high efficiency. We performed grid search cross-validation to find the best set of hyperparameters.

- learning_rate = 0.1 : Learning rate shrinks the contribution of each tree by learning_rate.

- n_estimators = 100 : The number of boosting stages to perform.

- max_depth = 3 : The maximum depth of the individual regression estimators.

- min_samples_split = 2 : The minimum number of samples required to split an internal node.

- min_samples_leaf = 1: The minimum number of samples required to be at a leaf node.

- Subsample = 1.0 : The fraction of samples to be used for fitting the individual base learners.

- random_state = 10 : Controls the random seed given to each Tree estimator at each boosting iteration.

VI. Model Evaluation

Out of all machine learning classifier models we used, Gradient Boosting Classifier performs the best in predicting house insecurity of the people in the dataset. The advantages of using Gradient Boosting Classifier are that it can optimize on different loss functions and provide tuning functions that make the model fit very flexible.

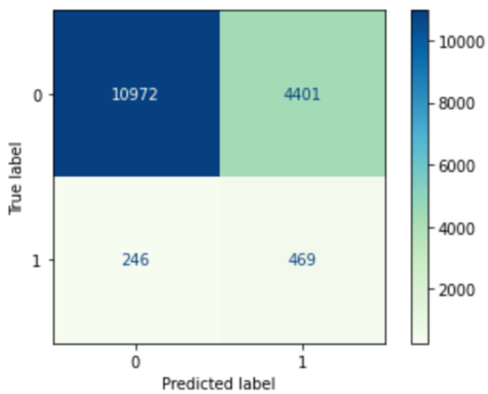
We created a Gradient Boosting Classifier model with an AUC score of 0.74, but it is crucial to examine additional model performance metrics in classification models to ensure that the created model can achieve the required outcomes. In order to analyze the performance of our model, we employed a confusion matrix.

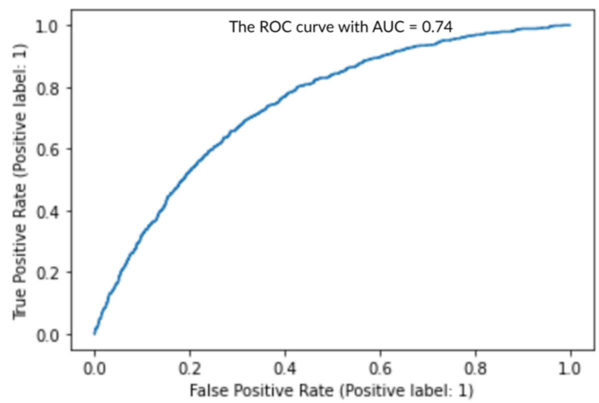
Sensitivity: A model's sensitivity refers to how well it can identify true positive cases. On dividing true positives by all positives, our model's sensitivity score is 72%, suggesting that it can correctly classify more than 2/3 of the members as non-Housing Insecurity.

Specificity: Specificity measures how well a model can identify true negative instances. The true negatives are divided by the total negatives to arrive at the calculation. Our model has a 65% specificity score, indicating that it can accurately identify 65% of the total members as Housing Insecurity.

Accuracy: Considering that we are utilizing an imbalanced dataset with a 4% HI – to non-HI ratio, our model's accuracy of 74% is not especially great.

Our model's performance in terms of specificity and sensitivity is neither exceptional nor very poor, which shows that the model is not overly biased in one way.


10972
4401
246
469
True label
Predicted label
0
1
0
1
2000
4000
6000
8000
10000


The ROC curve with AUC = 0.74
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

VII. Recommendations

Housing insecurity has been posted as one of the biggest social determinants of health. The health industry as a whole and Humana in particular, have been trying to sum efforts to address health by tackling housing issues. Such a complex and extended issue can't be addressed in only one way, but much good can derive from the synergy resulting from many efforts in different aspects of the problem.

People who suffer housing insecurity are exposed to health hazards such as trauma, chronic stress, and diseases directly derived from poor environmental conditions. [1]

Our research and analysis show that there are some characteristics of individuals that greatly relate to housing insecurity.

1. Mental Health

People with a history related to the following mental health issues showed a more direct relation with experiencing housing insecurity

- Mental and behavioral disorders due to psychoactive substance abuse in the past
- Claims related to psychosis, anxiety trauma, anxiety phobias, major or other depression, neurodevelopment disorders such as schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders.
- Use of prescriptions related to antipsychotic agents, depression agents, or other mental health related agents.

2. Drug/Substance abuse

People with a history of taking prescriptions related to substance abuse, a history of making claims related to substance abuse, or dependence on alcohol, tobacco, or other substances are more likely to experience housing insecurity in some of its forms.

3. Homeowner status

Members who are Renters or Probable Renters showed a greater chance of experiencing housing insecurity issues

4. Credit information

People with the following credit characteristics showed a greater correlation to housing insecurity.

- Non-Mortgage loans accounts with +60 days past due
- First Mortgage Accounts with +120 days past due or collections
- Large % balance to high mortgage credit

5. Geographical information

People live in counties in metro areas with larger population are more likely to have house insecurity issues. The geographical category implies two factors that increase people's house insecurity level: overcrowding and neighborhood safety. [Leopold, Cunningham, Posey, and Manuel, 2016] People in central city areas tend to live in condos and apartments that have limited space and there are cases of people sharing the same room in a unit. The negative impacts on quality of living due to overcrowding include increased physical contact, lack of sleep, lack of privacy and poor hygiene practices. These impacts will eventually take a toll on people's physical well-being and mental health, which contribute to their feeling of house insecurity. All statistics come to the conclusion that cities are more crime-prone to small towns. [Burbano, 2021] Since there are more people living in the crowded counties with metro service, there are more chances of people migrating around through affordable transportation and there are likely more crimes being committed. Living in a dangerous neighborhood is a form of house insecurity that can negatively affect the well-being of families. Especially for families with children, the stress of moving to a safer neighborhood with high-quality schools contributes to the sense of house insecurity. [Leopold, Cunningham, Posey, and Manuel, 2016]

6. Disability Status

Based on the training dataset, disabled people are more likely to experience house insecurity. A large number of disabled renters struggle to pay rent on time and are not confident in being able to pay next month's rent. [Lake, Novack, Ives-Rublee, 2021] People with more care needs experience higher rates of house insecurity.

7. Age

Based on the training dataset, people in their late 50s, 60s and early 70s are likely to experience house insecurity. This might be because people in their late 50s and 60s just retired and with a lack of income, they feel more financially insecure and housing insecure. People are more likely to develop physical and mental health issues at this age range and there are higher chances of dramatic changes in their life such as loss of family or caregivers, lack of social support due to retirement.

8. Race

Based on the training dataset, people who are black, American Indians and Hispanic are more likely to experience house insecurity. Due to the prolonged racial inequities in housing and in society, the result was expected and the disparity got worse due to pandemic. [Cai, Fremstad, Kalkat, 2021]

Our recommendation is to develop flags to detect these issues at an early stage when the member presents the first sign related to any of the above-mentioned characteristics. This will allow Humana to rapidly approach the person and try to provide the resources and health support needed to successfully overcome the corresponding issue. If not provided with the adequate resources or treatment in time, the person is much more vulnerable to continue worsening his condition and for it to negatively repercuss on his housing situation.

We suggest to collaborate with community based organizations and programs such as US Department of Housing and Urban Development to encourage natural surveillance within the community and promote mixed residential planning [Burbano, 2021], work with National Low Income Housing Coalition to come with rental assistance payment programs, free legal services and temporary financial supports, and reach out to American Psychiatric Association and World Health Organization to promote physical activity and mental care [Kiser, 2022], contact American Association of People with Disabilities to prohibit source of income discrimination, invest in accessible and affordable housing plans for disabled people, increase disabled renters protection. [Lake, Novack, Ives-Rublee, 2021]

[1] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://populationhealth.humana.com/wp-content/uploads/2020/06/Humana_HousingBrief_Final_External_version_2020.pdf

Improving Measures of House Insecurity: A Path Forward [Leopold, Cunningham, Posey, and Manuel, 2016] (2016, November 23) https://www.urban.org/sites/default/files/publication/101608/improving_measures_of_housing_insecurity.pdf

Why is Crime Higher in Cities Than in the Countryside? [Burbano, 2021] (2021, October 7) https://tomorrow.city/a/crime-in-cities

Fighting Housing Insecurity in the US. [Kiser, 2022] (2022, April 14) https://www.moneygeek.com/mortgage/resources/affordable-housing-and-assistance/

Recognizing and Addressing House Insecurity for Disabled Renters [Lake, Novack, Ives-Rublee, 2021] (2021, May 27) https://www.americanprogress.org/article/recognizing-addressing-housing-insecurity-disabled-renters/

Gradient Boosting in Classification: Not a Black Box Anymore! [Kurama, 2019] (2019) https://blog.paperspace.com/gradient-boosting-for-classification/#:~:text=Advantages%20of%20Gradient%20Boosting%20are,the%20function%20fit%20very%20flexible.

Housing Insecurity by Race and Place During the Pandemic [Cai, Fremstad, Kalkat, 2021] (2021, April 5) https://cepr.net/report/housing-insecurity-by-race-and-place-during-the-pandemic/