2022 HUMANA/MAYS

HEALTHCARE ANALYTICS CASE COMPETITION

IDENTIFICATION OF MEDICARE MEMBERS FACING HOUSING INSECURITY THROUGH PREDICTIVE MODELING AND RECOMMENDATIONS ON HOW TO HELP ACHIEVE THEIR BEST HEALTH

Table of Contents

1. Executive Summary

With the recent surge in housing prices after the pandemic, a wide range of problems related to housing insecurity have become a priority concern in the healthcare industry. The U.S. Government has reported that 580,000 people suffered from homelessness in the U.S. in 2020 and for those with home insecurities, 53% had severe behavioral disorder including substance use disorder or depression, and 17% had medical problems related to asthma or hypertension.

This study focused on helping Humana, as a leading healthcare insurance and services provider, predict members who are likely to face housing insecurity issues and reduce the risk of related health problems by diagnosing and preventing them at an early stage.

Our goal was to develop a model that predicts the probability of a member experiencing housing insecurity in the future. Through data analysis and predictive modeling, we generate a model that can identify members most likely to be struggling with housing insecurity issues. This model, which has a 0.7573 ROC-AUC score, is also robust in fairness with regard to gender and race, as evidenced by the high disparity score of 0.99441.

Based on the member information segment, we derived actionable insights to propose potential distinctive solutions. Our solutions consist of four different segments based on the top features of our model. We suggest that providing the following preventive solutions to targeted members at Humana can be potentially cost effective in the long term.

- Age segment: Investment in housing and rent to young members with low income at affordable rates, and in return, Humana can save from reduced medical claims caused by housing insecurities and acquire more young members to solve member age imbalance.
- Medical segment: Regular check-ups on members at high risk, characterized as having high interaction count, diabetes claims, or high behavioral medical cost, to provide preventive cares
- Regional environment segment: Improving environmental conditions for members in poor housing conditions or high polluted areas to prevent respiratory diseases

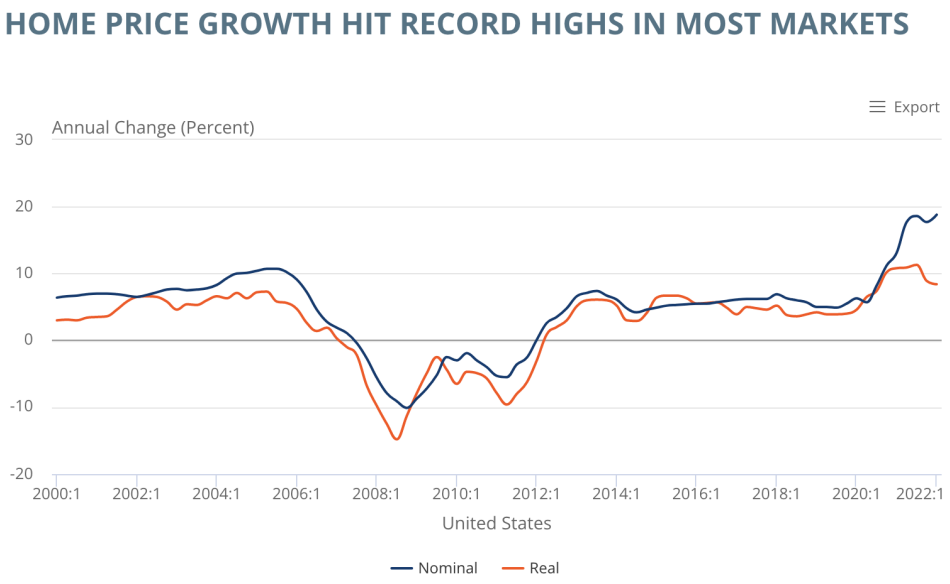
Further improvements can be made to the model with suggested data collection strategies.

2. Introduction

2.1. Case background - understanding of competition objective and problem statement

With the recent surge in housing prices after the pandemic, a wide range of problems related to housing insecurity have become a priority concern in the healthcare industry. Housing insecurity is defined as a general term encompassing several dimensions of housing problems including but not limited to difficulties in paying rent, frequent moving, overcrowding, or high rent-to-income ratio[1]. People who are housing insecure have higher chances of getting exposed to serious medical and behavioral health problems, which may result in increase in medical expenses.

According to statistics from the U.S. Department of Health and Human Services in 2020, about 580,000 people experienced homelessness in the U.S., and among those, 53% had substance use disorder, 35% had major depression, and 17% had asthma or hypertension. The more serious problems are also related to increased risk of premature death[2].



[Figure 1: Housing price growth in US[3]]

As a historic number of people have reported that they have lost their jobs or houses after the pandemic and even with added influence of economic inflation forecast, the housing insecurity problems are predicted to continue rising in the future. In the need of solutions to these problems, housing subsidies administered by the federal government are designed to provide financial assistance to help people with low income, but the statistics show that only 26% of people in need received these supports due to limited funds and long waitlist[2]. For Humana, as a leading

healthcare insurance and services provider, it is beneficial to find solutions that can identify members with housing insecurity and reduce the risk of related health problems by diagnosing and preventing them at an early stage.

On the grounds of this, we have constructed a predictive model that identifies members who are most likely suffering from housing insecurity and come up with suggestions on how to prevent health risks posed by housing insecurity. Through this competition, we hope that our model contributes to Humana in proposing potential solutions on how to provide these members in need with more intensive and preventive care.

2.2. Competition goal metric definition

For this competition, the prediction model will output a probability that a particular member is suffering from housing insecurity and based on this probability, return a binary indicator where 0 is negative prediction and 1 is positive prediction in housing insecurity. The model will be judged based on both accuracy and fairness by measuring the following metrics.

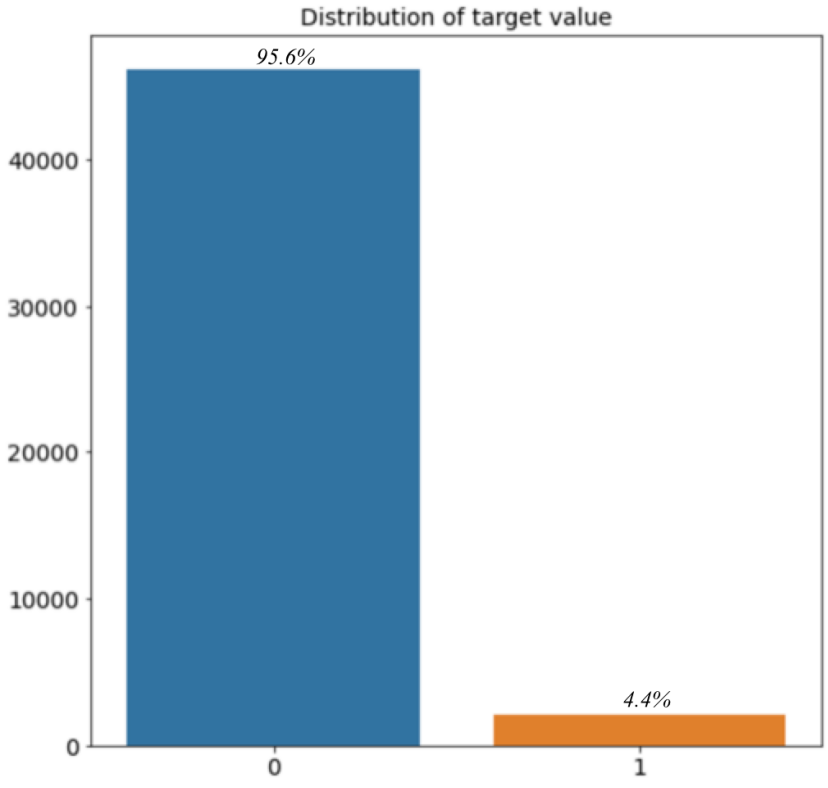
For accuracy, the model will be evaluated using ROC-AUC score[4], which calculates the area under the receiver operating characteristic curve. The ROC curve plots a true positive rate and false positive rate, and we aim to increase the true positive rate but decrease the false negative rate. The AUC score is known as an accuracy measure for binary or multiclass classification models which fit this competition. The higher the AUC, the better the performance of the model at classifying between the positive and negative labels. In this competition, a higher AUC indicates that the model performs better in predicting the members with housing insecurities.

For fairness, a disparity score will be used to measure the possible bias in the model. Bias is inherently present in the real-world data in different forms such as historical stereotypes, representation, or measurement. We cannot remove the bias from the data directly, but we can take proper measures to avoid them in modeling. A disparity score measures the bias by comparing the ratio between two groups that receive a positive prediction. The best practice to address the bias issue is to calibrate for each group. As Humana is committed to supporting inclusive AI for protected classes under the category of race, sex, age, low-income status, and disability status, the model should generate a prediction leading to fair and equitable outcomes.

3. Exploratory Data Analysis

3.1. Data overview

To build a predictive model that identifies members who struggle with house insecurity, we were provided with two separate datasets: a training dataset and a holdout dataset. The training data consists of 48,300 records with 880 features and a target column that flags a member as having housing insecurity. The holdout data consists of 12,220 rows with the same number of features as the training dataset but with the target variable held out. The training dataset is highly imbalanced with the 4% of the members in the dataset being flagged as struggling with housing insecurity (binary class of 1) and 96% not being flagged. Below table and barplot show a detailed summary of the provided datasets and the proportion of our target value.



[Figure 2: Target value distribution]

Dataset	Number of Rows	Number of Features	Target column
Training dataset	48,300	880	1
Testing dataset	12,220	880	Not applicable

[Table 1: Overall dataset information]

Before moving onto data preparation and cleaning, we first concatenated the two datasets so that data transformation and feature engineering can be done in a consistent manner in both sets of data. After combining the two datasets, we confirmed that there was no data leakage between training and test data by dropping all duplicate values. After confirming that the data is free of any issue and preparing it for modeling, the combined dataset was divided back into separate training and test sets prior to training our model.

3.2. Missing values

First, we began by searching for missing values in each column in order to decide whether to drop these columns from our dataset or impute the missing rows with another value. Out of 880 total feature columns, 267 columns have at least one missing value. Below table shows only the columns that have the highest percentage of the data missing. However, a full list of columns with at least one null value along with the count and percentage of rows per column can be found in Appendix A.

	Missing count	% Missing value
cms_risk_adj_payment_rate_b_amt	60494	0.999570
cms_tot_partd_payment_amt	57585	0.951504
credit_hh_agencyfirstmtg_new	56754	0.937773
credit_bal_autobank_new	56691	0.936732
credit_num_autobank_new	56678	0.936517
credit_num_nonmtgcredit_60dpd	56675	0.936467
credit_prcnt_mtgcredit	56664	0.936286
credit_bal_bankcard_severederog	56646	0.935988
credit_bal_consumerfinance_new	56617	0.935509
credit_hh_autobank	56607	0.935344
credit_bal_nonmtgcredit_60dpd	56592	0.935096
credit_num_1stmtg_collections	54460	0.899868

[Table 2: Top columns with missing values]

For the columns that have greater than 95% of its data missing, our plan is to drop them from our dataset as these columns do not contain enough data from which we can determine what value to impute the missing values with.

3.3. Data type inconsistency

Next, we looked for any inconsistencies in the data types within each column. For example, there are null values within the original dataset that are indicated in the format of type string, "*", and a column that has a combination of numerical and categorical values, which needed to be cleaned up for further analysis. In addition, columns ending with *'_ind'* or *'_cd'* contained categorical values but were identified as numerical columns, so they need to be converted into categorical columns to prevent possible misunderstanding in correlations and other aggregated calculations.

3.4. Feature analysis by subcategories

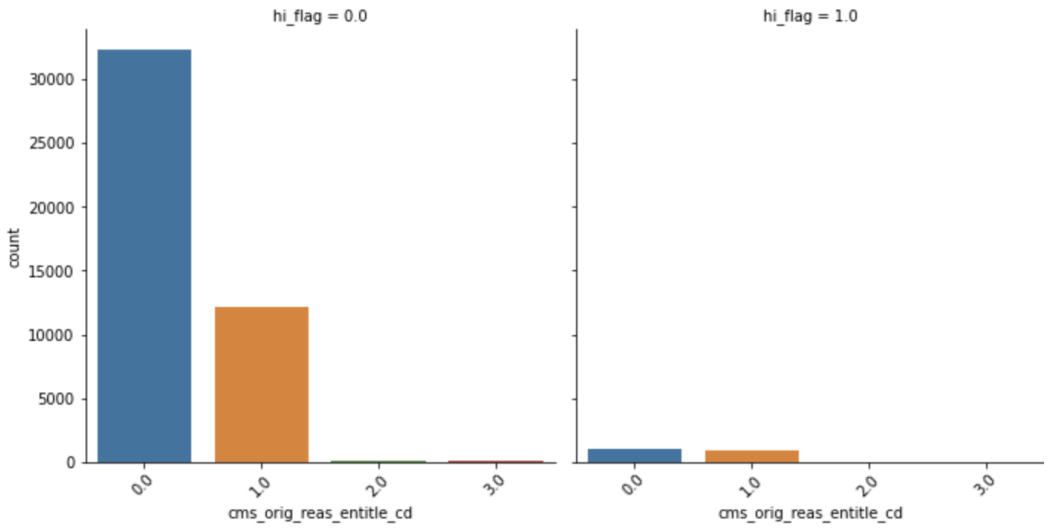
For further analysis, we also divided the features into subgroups by the prefix, suffix, specific keywords in the name of each column, such as *bh, cci, cmsd1, ind, rx, rev,* and *total,* and their combinations. The level of understanding that resulted from this exploratory data analysis allowed us to engineer several new features for our predictive model, which will be discussed in the next section. Below are some highlights for each subgroups:

prefix	details	measurement
atlas	macro factors based on population	percent, count
bh	behavioral health claims	count, cost
cci	health claims based on Charlson Comorbidity Index	count
cms	claims in medicare & mecicaid services	binary/categorical code, amount, score
cmsd1	claims based on CMS diagnosis code level 1	count
cmsd2	claims based on CMS diagnosis code level 2	count, binary indicator
cnt	count per month of member interactions	count
cons	consumer information by individual level	binary/categorical code, percent, score
med	days since last claim for non-behavioral health claims	count
rev	claim lines per month for a revenue code	count
rwjf	macro factors by Robert Wood Johnson Foundation	average, ratio, count
rx	prescriptions per month by disease	count, cost
total	overall claims/cost information	days, count, cost

[Table 3: Top columns with missing values]

3.5. Distribution analysis by categorical features of members

Since another one of our goals is to come up with a targeted solution for different segments of members, we checked the distribution of housing insecureness by each member's various status, such as the reason code for medicare entry, type of risk adjustment factors, geographic information, race and sex. While most of the features show similar distributions for both hi_flag = 0 and 1, we found that the reason code for medicare entry showed significant differences in distributions on the target value. The varying distributions among different subgroups by their housing insecurity classification status can be found in the plots in the next page.



[Figure 3: Target value distribution by reason code]

3.6. Hypothesis testing

Next, hypothesis testing was conducted to check if a given feature has statistical significance on the target value or other features that have relatively high impact on classification of the target value. First hypothesis testing was to check the distribution difference of *hi_flag* by *cms_orig_reas_entitle_cd*. We have divided the *hi_flag* distribution by the original reason for entry into Medicare in order to conduct a preliminary hypothesis test to determine whether the distributions were significantly different among different groups.

- Null Hypothesis, H_0 : *hi_flag* distribution is constant regardless of the original entry reason

- Alternative Hypothesis, H_A : *hi_flag* classification differs by the original reason for entry into Medicare

Kruskal Wallis Test was conducted to compare the distribution of *hi_flag* across the original reason categories. p-value was essentially 0 (p-value=8.937420360232403e-83). As a result, we reject the null hypothesis at the α=0.005 significance level and conclude that *hi_flag* distribution differs by the original reason for entry into Medicare.

Another hypothesis testing was performed to check for differences in the distribution of *atlas_age65andolderpct2010* by *rucc_category*. Similar data preparation was conducted as above, with the hypothesis set up as below.

- Null Hypothesis, H_0 : *atlas_age65andolderpct2010* distribution is constant regardless of *rucc_category*
- Alternative Hypothesis, H_A : *atlas_age65andolderpct2010* distribution differs by r*ucc_category*

As this testing was also comparing distribution by different categories, Kruskal Wallis Test was conducted, which resulted in a p-value of 0. As a result, we reject the null hypothesis at the α=0.005 significance level and conclude that the *atlas_age65andolderpct2010* distribution differs by *rucc_category*.

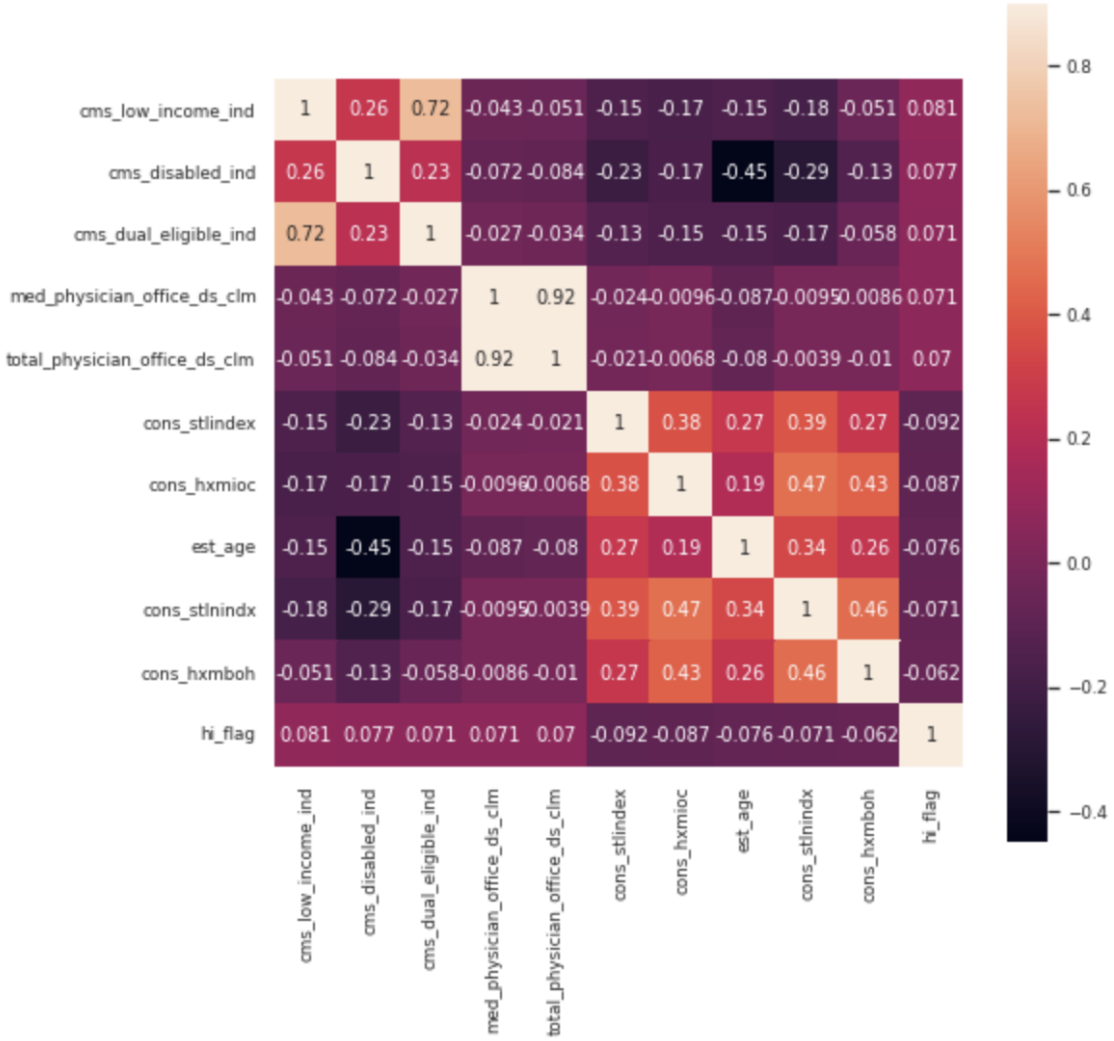
Test topic	Test	Statistic	P-value
hi_flag distribution on cms_orig_reas_entitle_cd	Kruskal-Wallis	383.351	8.93E-83 ≈ 0
atlas_age65andolderpct2010 distribution on rucc_category	Kruskal-Wallis	8,520.065	0.0 ≈ 0

[Table 4: Hypothesis testing results]

These hypothesis tests that we conducted were a means to get a sense of whether a feature would serve as meaningful features in distinguishing members who are highly likely to be facing housing insecurities from those who are not.

3.7. Correlation analysis

We also performed correlation analysis between the numeral features with the target value. The top positive/negative features were as below.



[Figure 4: Heatmap of high correlated features on target column]

Based on the correlation plot above, we can see that some indicators such as the low income indicator and disabled indicator have an impact on housing insecurities. Also, we can see that estimated age points to an opposite direction on having housing insecurities, which generally makes sense as a younger person has not had a chance to work long enough to become financially stable yet. Thus, he or she is not able to afford housing, which can contribute to the probability of facing housing insecurity.

4. Feature Engineering

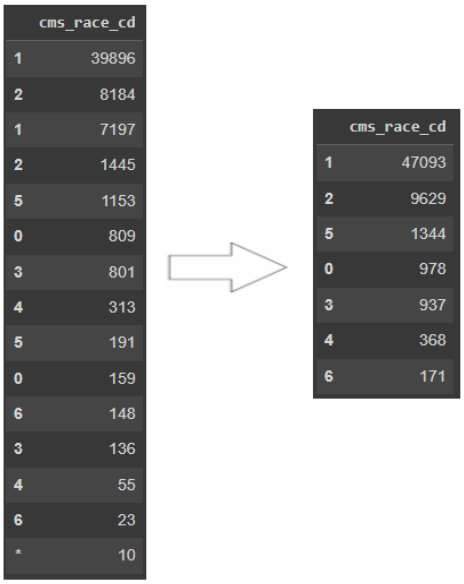
After gaining an in-depth understanding of the dataset through EDA, we moved onto cleaning the datasets, which would address the issues that we found in our original datasets using proper imputation methods, removing features, and creating new features from existing features. After finishing feature engineering, we had a total of 48,300 rows with 343 columns. Note that we've included all the outliers, as outliers also contain important information of a member and need to be included in the analysis. Moreover, there weren't any outliers that seemed highly unusual, which we would have otherwise interpreted as incorrect data and either dropped from our dataset or imputed with another value.

4.1. Dataset clean-up

As shown in the EDA part, we first cleaned the dataset using the following methods:

- Replacing symbols with null
- Matching data type to be uniform within each column
- Converting data types to reflect the features correctly

For example, we replaced the symbol, "*"', with either NaN, another string, or a number in a format of a string, while also taking into consideration whether the features that are labeled as being numerical were, in reality, numerical or whether they were given a numerical label in means of a categorical feature. For example, for the *cms_race_cd* column, which uses a numerical code to indicate a member's race, we interpreted "*" as *unknown* and assigned a corresponding preexisting label of 0. As for the duplicate race indicator values, we took the sum and consolidated them into a single code for each race type.



[Figure 5: Example of matching data type]

4.2. Missing value imputation

Per our initial exploratory data analysis, the two columns with over 95% missing values were removed from the dataset. This approach resulted in a removal of two columns.

For the numerical columns that have missing values but have some keywords shared with other columns, we selected the column as a reference and replaced the null values with the subcategorized median per group based on the referenced column.

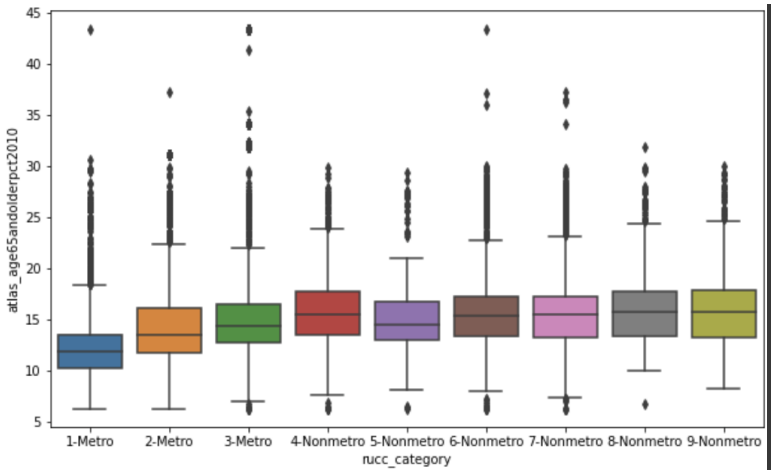
For categorical columns, the null values are merged into ‘U’ (Unknown) or ‘OTH’ (Others) if these labels already exist in the data, and for features without such existing unknown labels, we created a new label for the null values. Similar idea that was used for numerical columns was used to replace null values in categorical columns that share the same keywords. For example, to fill the null values in the spoken language column, we’ve referenced the column with race.

4.3. New feature generation

In order to further improve our model accuracy, we created new features based on the values of existing columns.

4.3.1. By geographic segmentation

First, we grouped by the geographic region information column, *rucc_category,* to compute the aggregated mean and standard deviation for the *atlas_age65andolderpct2010* column, the percent of population 65 or older, which consistently made it on our list of top high importance features in modeling, and the number of members who live in each geographic region. In the boxplot below, we can observe notable differences in the distribution of the percent of population 65 or older for each group, which allows us to extract new, potentially useful features based on these varying distributions.



rucc_category	mean	std	count
1-Metro	12.283826	3.463695	26380.0
2-Metro	14.451612	4.225405	16181.0
3-Metro	15.730000	5.740831	6622.0
4-Nonmetro	15.716145	3.259394	3375.0
5-Nonmetro	14.745877	3.058238	1060.0
6-Nonmetro	15.575419	3.378335	4095.0
7-Nonmetro	15.817272	3.896012	1796.0
8-Nonmetro	16.108336	4.032870	571.0
9-Nonmetro	16.288636	4.301741	440.0

[Figure 6: Distribution of age 65 or older percentile by geographical segmentation]

By joining the below descriptive statistics table with our original dataset, we are able to gain information on the average percentage of population 65 or older for the region that each member lives in, the standard deviation of the distribution for the average percentage of population 65 or older, as well as the total number of members living in that region.

Using a similar approach, we calculated the percentage of each race in the region that each member lives in by grouping by the *rucc_category* and *cms_race_cd* column, and then dividing the total count of each race by the total count of members in each region. These engineered

features that were created using this method provide us information about the members at the level of geographical segmentation that was not evident from simply looking at the already existing features.

4.3.2. By prefix

As we mentioned in the EDA section above, we observed features with several prefixes, such as *bh, cci, cmsd1, ind, rx, rev,* and *total,* and their combinations. We figured that these columns are mostly having 0's or a single value but overly segmented, thus do not effectively offer underlying insights to the target. Therefore, we used a similar approach as we did in geographic features above and grouped together for columns having the same prefix to generate new features having aggregated sum and mean values per prefix. For example, all columns that are related to count of visits for behavioral health were combined together, and a new column was created having sum of all the behavioral health information per member. This approach created 66 new features.

Additionally, using the created new features of sum and mean above, we created another feature that categorizes the sum and mean values to four different categories of low, medium, high, and very high to further generalize the features. Since the original individual feature is overly segmented, it might overfit to the training data and prevent the model from providing generalized prediction for the unseen data. This approach created 66 additional features to the model.

4.3.3. By keywords: medical and behavioral health conditions

Based on the statistics that people who experienced homelessness show higher chances to suffer from certain diseases such as substance use disorder, depression, asthma, or hypertension[2], the columns related to diseases above were grouped separately by disease to produce new features indicating the likelihood of having these diseases. Per health condition category, the individual column shows that most values are focused to a single value or 0, so the features can be more generalized by looking at the aggregated values. Therefore, the sum of these sub-categorized columns per disease were generated and these columns are given binary indicators that classify 0 for members who do not have any claims related to the corresponding disease and 1 for members

with past claims of the corresponding disease. This approach created additional 4 features to the model.

4.3.4. Other methodologies

Other new features were also generated by using the following strategies. For features related to personal income, we already have the Census Income Percentile but a direct comparison of individual income percentile may not give us a proper diagnosis as the income level can be highly region-dependent. Therefore, we divided the Census Income Percentile by scaled median house income per group in region feature, *rucc_category*, to see if the person still has high or low income compared to the median income of the area. Based on this idea, a new column is created that if this ratio is greater than 1, the specific member was categorized as a high income level.

Also, some columns such as total number of occupied housing units may not correctly represent the percentile of occupancy because occupied housing units can be highly dependent on population. For these columns, we divided them by the population to scale the numbers that can better reflect the given data.

4.4. Feature selection

Selecting appropriate features is essential for enhancing model performance. We dropped the features that were overly segmented in our original dataset after they were used to generate new features by aggregating by grouped features. Also, for the features that contain only one unique value, we dropped them as they do not help the prediction but only increase tasks to the model. Therefore, the resulting features were reduced from initial 881 columns to 343 columns.

4.5. Feature scaling and encoding

As a step of feature engineering, we used a standard scaler to normalize numerical features. For categorical features, we used label-encoding rather than one-hot encoding, as both methods show similar performance, but one-hot encoding might increase the sparsity of the dataset.

5. Modeling

5.1. Train test split

We splitted the training dataset into training and validation by 80:20. This split ratio was selected as we were considering 5-fold as a cross validation.

5.2. Model selection

We selected LightGBM[5] as a final model. As the problem was to solve binary classification problems with non-linear features, we expected that tree-based boosting methodology would perform better than other models. Also, among the tree-based boosting models, LightGBM was the most efficient recipe to train and test the performance as this model is one of the fastest boosting models in terms of computation time, while maintaining its prediction power.

We also conducted some empirical analysis by comparing several different models such as Logistic Regression, SVM (Supportive Vector Machine), Decision Tree, Random Forest, XGBoost and CatBoost. Among all models, we figured out that LightGBM performed the best both in terms of performance and computation time.

Model	Valiation_Score
LightGBM	0.74467
XGBoost	0.74040
CatBoost	0.73819
Logistic Regression	0.73019
Random Forest	0.72599
AdaBoost	0.72575
Decision Tree	0.67761
K-neighbors Classifier	0.53355

*Metric: ROC-AUC score

[Table 5: Model performance comparison]

5.3. Hyper parameter tuning

To improve prediction power, we conducted hyperparameter tuning firstly by using Bayesian search[6]. This method is well-known and one of the most effective hyperparameter searching methods as we only need to set some parameter range and searching results return the local optimal parameters within the ranges. As this method might not be giving the global optimal parameters, we ran several bayesian searches, and also manually confirmed the parameters based on the local optimal parameters we found. Some key parameters and their values are as follows.

n_estimators	learning rate	num_leaves	subsample	max_depth	lambda_l1	lambda_l2	colsample bytree
5000	0.005	32	0.8	12	0.8	10	0.5

[Table 6: Hyper parameter used in our final model]

5.4. Cross validation with K-folding

As the final step of modeling, OOF[7] (Out Of Folding) method was used as a cross-validation methodology. 5 folds were used as we intended to split the train dataset into train/validation into a 80:20 ratio. By conducting the OOF method, 5 different iterations were averaged to get a prediction on the test dataset.

6. Model Result and Analysis

Based on the prepared dataset and model, we derived the model result and evaluated its performance. We also checked if the model was biased, and analyzed some important findings from which we based our recommendation, which will be presented in the next section.

6.1. Model performance

As the model is evaluated by ROC-AUC score, we got AUC score as 0.75776 for the validation score, and 0.75736 for test score. This result shows that the model has robust predictive power both on validation and test set.

Model	Valiation_Score	Test Score
LightGBM-Hyper parameter tuned	0.75776	0.75736
LightGBM	0.74467	

*Metric: ROC-AUC score

[Table 7: Final model performance]

6.2. Model bias analysis

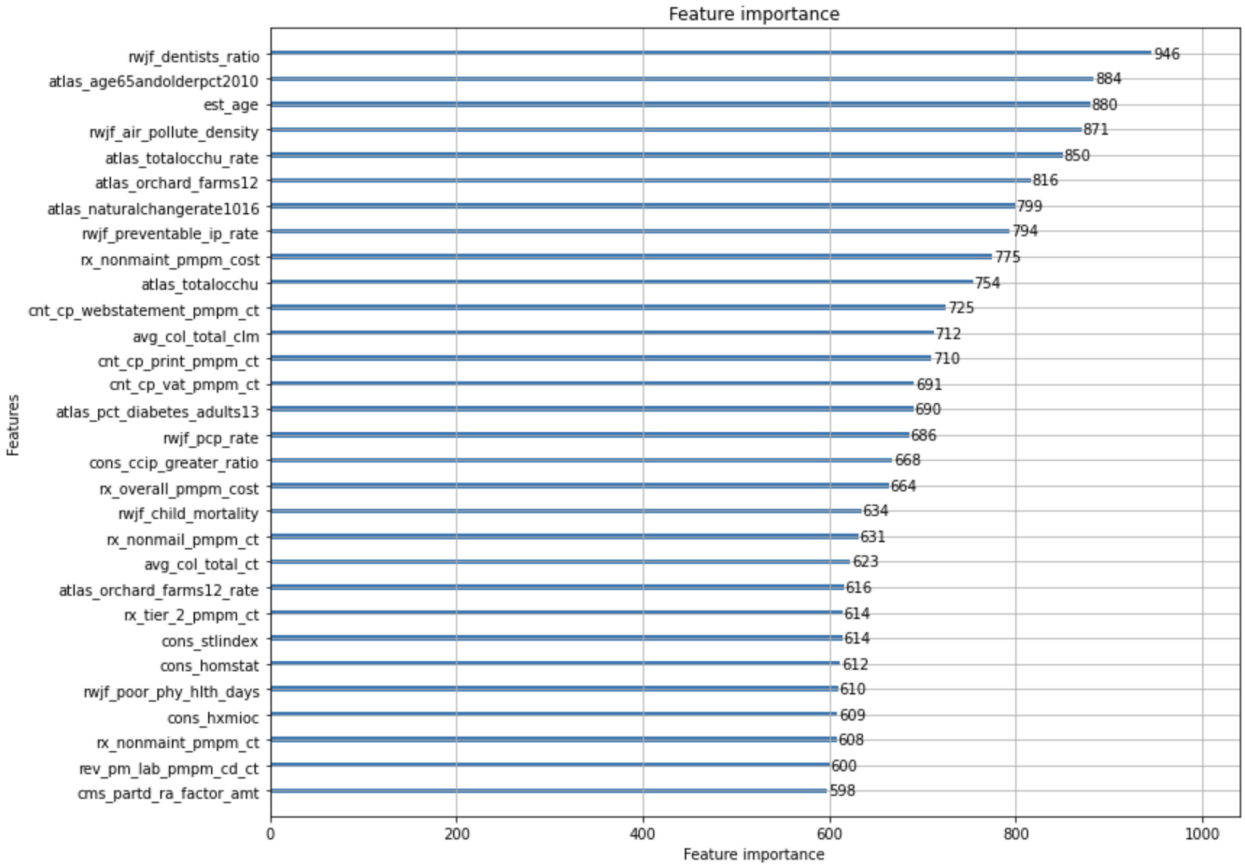
We also evaluated the model's performance by race and sex, as disparity score and corresponding weighting will be calculated utilizing those two factors. Some categories showed low performance, but a possible explanation for this is the low volume of training dataset. As shown below, our model shows robust and unbiased performance by race and sex.

Segmentation	Sub category	Number of data (training set)	ROC-AUC (validation set)
Sex	Female	29100	0.75525
	Male	19200	0.75573
Race	White	37549	0.75206
	Black	7706	0.76345
	Hispanic	1068	0.66791
	Unknown	788	0.86215
	Other	759	0.70647
	Asian, Asian American, or Pacific Islander	298	0.61111
	American Indian or Alaska Native	132	0.85185

[Table 8: Model bias analysis]

6.3. Model analysis

The feature importance plot generated by LightGBM allows us to visualize the magnitude of impact that each of selected features has on classifying our target variable. Below plot displays the features that have the overall highest importance in predicting the probability of housing insecurity issues. While there are some macro features to be considered, we can observe that some of the features that we created as a part of feature engineering such as *avg_col_total_clm, avg_col_total_ct, avg_col_total_cost,* and *avg_col_med_clm* carry a significant level of importance in our model performance.



[Figure 7: Top 30 features by feature importance]

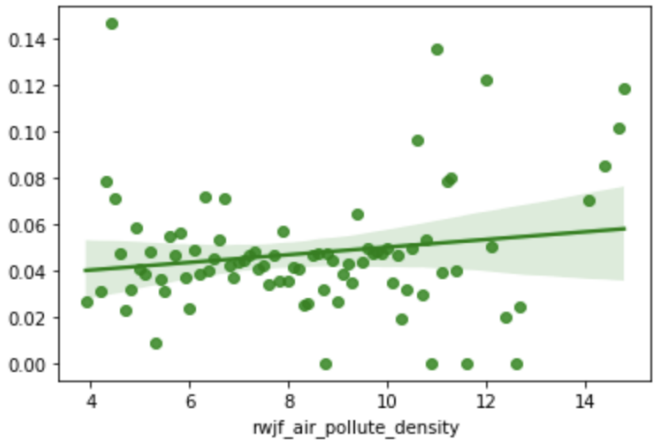
Among top features, we categorized some top features by components of housing insecurity. Categorized top features are as follows:

- Clinical Care: Number of dentists, primary care physicians
- Age: Age over 65 ratio, estimate age
- Environment situation: Air pollution density
- Member interactions: member interaction via web statement, print, and virtual assistant
- Financial status: income level compared to the region's median, short-term loan status
- Housing status: rate of occupied housing units, homeowner status

Also, below are more detailed potential explanations of top 5 features.

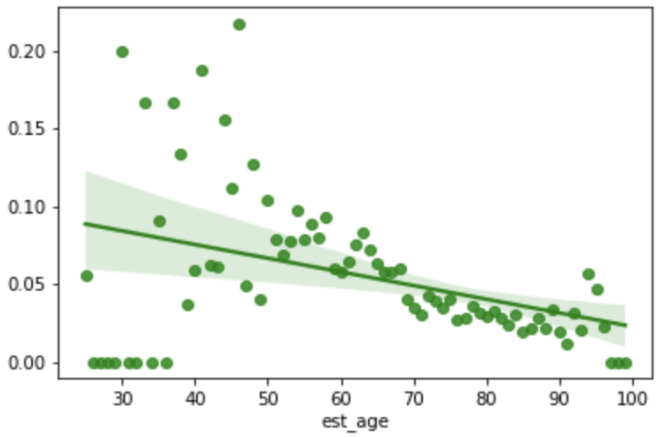
- *rwjf_dentists_ratio*: While this feature does not show high correlation with its target value, it gives evidence that clinical care is highly related to housing insecurities.

- *atlas_age65andolderpct2010*: As this feature has some negative correlation with the target value, it shows that the younger members are more likely to have housing insecurity issues.
- *air_pollute_density*: This feature gives an idea that the living environment is an important factor for housing insecurities. The more the air pollution density is, the more the members are likely to have housing insecurities.



[Figure 8: Correlation plot: *hi_flag* vs *rwjf_air_pollute_density*]

- *Atlas_totalocchu_rate*: This feature gives some sense that housing occupancy makes people more stabilized in terms of housing.
- *est_age*: As mentioned in the age 65 population ratio, this feature strengthens the tendency that as you are younger, it is more likely to get housing insecurity. This also indicates that we need to set a strategy/recommendation based on its age level.



[Figure 9: Correlation plot: *hi_flag* vs *est_age*]

7. Proposed solutions

With ongoing housing investment and funds in Humana[8], finding an optimal way to distribute the possible funds is an essential starting point. Based on our predictive model, we created the following baseline recommendations per sector using the top impacting features from the model to efficiently distribute the existing and future resources in Humana.

7.1 Age segment

7.1.1 Support for younger generations: direct housing support

As one of the top features in our model, *est_age* shows that as younger the member is, the higher chance of facing housing insecurity. The younger generations have a higher chance of being financially unstable, possibly due to lack of work experiences which may lead to lower income. Also, the continuing increase in the housing market hinders younger generations from acquiring housing. As a solution, investment in housing and temporarily renting the units to the members at affordable rates can relieve the issue directly. In addition to this finding, it is clear in the data that the median of *est_age* is 72.0 which means that the percentage of members with younger ages is extremely low. As a marketing effect, providing temporary housing to younger generations could also resolve the issue with age imbalance in member ratio and it can increase the company's revenue in return[9] .

7.1.2 Support for older generations: connect with government support resources

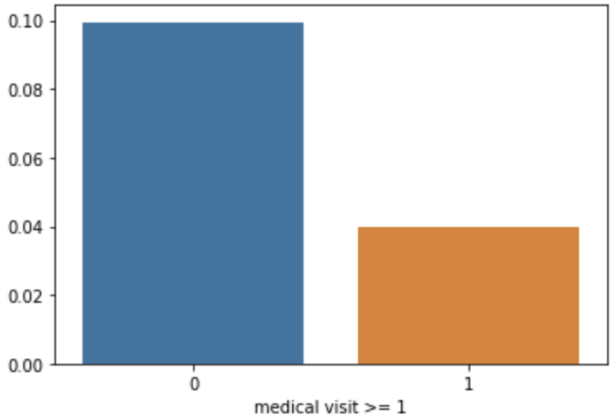
Although the model predicts the younger generation is in higher needs, we should also consider the older generation of 65 years or above since they are the majority of members at Humana. Based on our investigation of the income data grouped by age, the income level of older generations is higher than the younger generations. Therefore, it seems that the support for older generations is not necessarily direct financial support, but rather utilizing online resources for retirees or senior housing. For those of age of 65 or above but with relatively low income, the possible solutions can include connecting with government support on behalf of older generations since the older generations may experience difficulties in utilizing online resources and not be able to find eligible support.

7.2 Medical segment

Although providing direct solutions to housing insecurities would be the most effective way to resolve the issue, we have to face the issue of having limited resources. As a secondary step to the primary support above, maintaining health conditions could be the second most effective way to prevent health related issues and thus reduce the medical cost of the members with housing insecurities.

7.2.1. Free health check-up for first-time patients

Based on the data provided, the probability of having housing insecurity when the member has never visited any medical services is 2.5 times higher than the probability of having housing insecurity compared to the member who has visited medical service at least once.



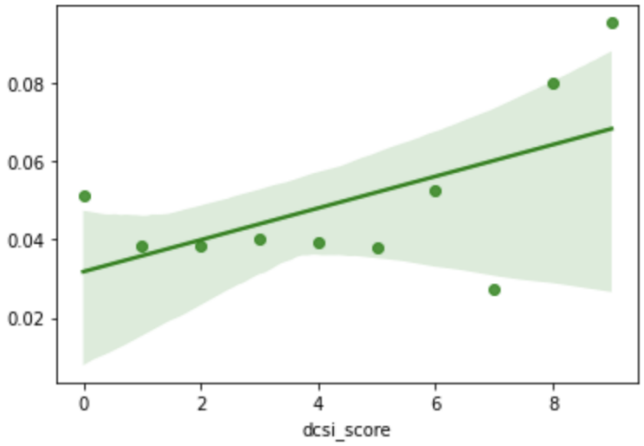
[Figure 10: *hi_flag* occurrence by number of medical visits]

Finding the member's medical issue at an early stage is as important as maintaining a good health condition. Therefore, it would be helpful if we can provide and encourage them to utilize free health check-ups for those members who have never received a health check-up.

7.2.2 Health Maintenance / Improvement Incentive Program

Managing and maintaining good health conditions of members are the primary concern of healthcare service providers. The top feature plot supports the connection with diabetes rates and

housing insecurities. For example, this regression plot indicates positive correlation in *dcsi_score* (Diabetes Complication and Severity Index score) and *hi_flag* probability.



[Figure 11: Correlation plot: *hi_flag* vs *dcsi_score*]

More importantly, as *dcsi_score* is based on the claim, there are some potential issues for non-reported diabetes. It is possible that some members live without knowing their underlying health conditions until after a proper diagnosis by a doctor. Therefore, it is important for Humana members to seek preventive healthcare. Practical solutions for preventing the related diseases can include following:

- Giving incentives for our members to get regular check-ups
- Giving discounts or additional discounts to members whose health shows improvement from previous year
- Providing members with self check up health kits that can be used at home
- Providing proper nutrition support and education program to members

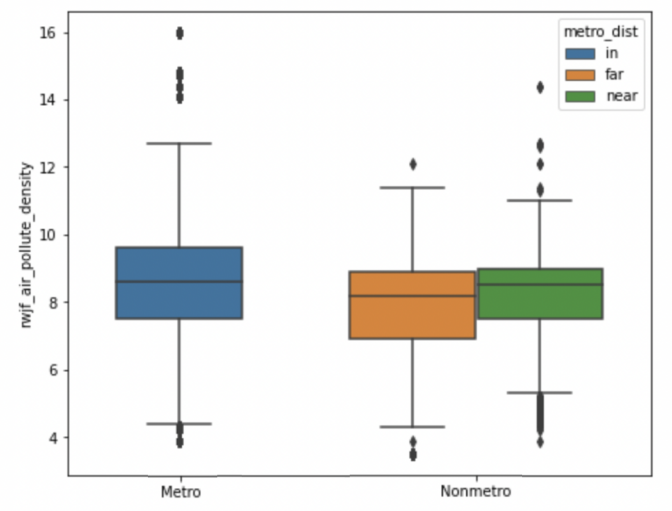
This is a viable program to offer for Humana members since many companies are already offering the same type of health incentive programs for their employees. The type of incentives given to the members can potentially be linked to the reimbursement cost for improving housing quality, which is discussed in detail under section 7.2.4. Motivating the members to increase self-care and awareness of their health conditions can help prevent more serious health issues in the future because left unchecked, disease and illness can result in poor health, chronic problems, and even death

7.2.3. Psychological Counseling

Housing instability might cause psychological health issues as well. A study[10] has shown that substance use disorders and severe depressions are often observed from the people experiencing homelessness so we should not underestimate the mental health of Humana members. Based on the *rx_bh_pmpm_cost* (cost per month of prescriptions related to behavioral health drugs) feature, members who have housing insecurity spend approximately twice more compared to members without housing insecurities. As mental health can get worse in a short period of time and lead to more severe consequences, enhancement in psychological counseling may help members to overcome or prevent major depressions or behavioral disorders resulting from housing insecurity. Moreover, as a long term effect in return, members might eventually not need further help in behavioral health components, and this would be one of the desirable ways to enhance the medical status of members at Humana.

7.3 Regional environment segment

As shown in the below plot, we can observe the differences in the level of air pollution density for each geographic region and its relative proximity separated by either metro area or non-metro area.



[Figure 12: Air pollution density by regional segment]

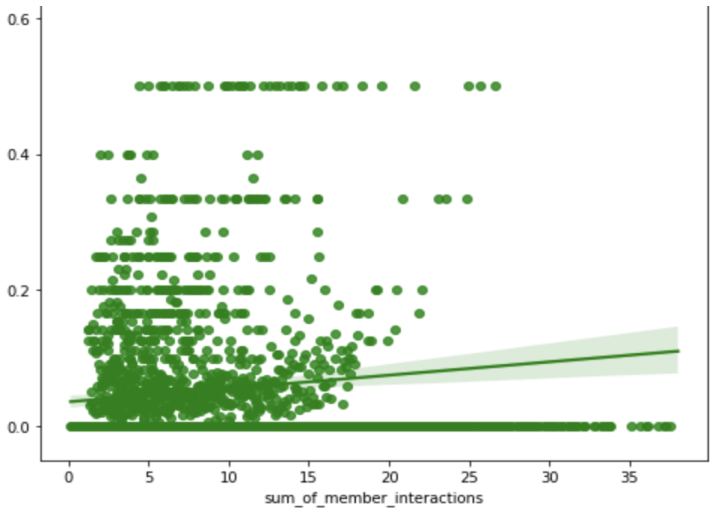
The effects of poor housing conditions - for example, a lack of adequate ventilation - is an influential risk factor on occupants' respiratory health[11], and levels of pollutants found indoors can be even higher than the outdoor levels if a home is not properly ventilated[12]. Living in a housing that has poor conditions can also lead to risked health conditions that are related to housing insecurities. Therefore, providing the members with reimbursement on expenses that are used to repair or reconstruct their homes is a potential solution to helping those members who are likely to be facing housing insecurity as we saw that pollution density level had a positive correlation with the probability of a member facing housing insecurity . Studies have also shown that home design and structure are found to significantly influence housing quality and may affect both mental and physical health[13]. The addition of this housing condition improvement program can aid in providing better healthcare for the members.

7.4 Internal developments

While above solutions are appropriate to be considered, we need to make sure that we are targeting the right members. Also, we may want to use important features to presensor members that are at risk of housing insecurities. Based on the sources that we have, we are proposing some solutions below that can be further developed internally at Humana.

7.4.1 Managing members with high interactions

Based on our research, the more the member has the interaction (via email, mail, call, etc) with Humana, the more the member is likely to have housing insecurity issues. Therefore, it is important to keep in touch with members that have high interactions with Humana and on top of answering claim related questions, we recommend asking them regularly if they are facing any housing related problems. In case they are facing any housing related problems, providing or utilizing internal legal help at Humana might resolve their issues at an early stage and prevent possible evictions.



[Figure 13: Correlation plot: *total number of member interactions* vs *hi_flag*]

7.4.2 Managing members with frequent moving

In addition to the support above, we can consider another issue related to frequent moving which is another factor of measuring housing insecurities. Frequent moving can be a potential indication of eviction[2], so keeping track of address change of members would give us additional metrics to the likelihood of possible housing insecurities. Even if the frequent moving was not caused by eviction, frequent moving itself explains that the member is likely experiencing problems with stabilization. For this type of members with frequent moving history, a community-wise support can be helpful for them to adjust into the current living environment, for example, providing local community based programs to motivate and build bonds between neighbors.

7.4.3. Further data collection

To enhance the prediction power of the presented model and to give more concrete evidence of the solutions shown above, we propose some of the methods that would make the Humana dataset more fruitful. Some useful features that might be relevant for making a better prediction in identifying members with housing insecurity issues include the number of times a member has moved in a given period of time (for example, in the past X years or months), which can be either tracked by an address change in the system or through a survey. Also, as there were too

many 0's on count or cost features, it would be better to create a feature that categorizes features into zero or non-zeros. Including survey results from the members would also be beneficial. For example, if we can have some survey results stating that the member is feeling insecurities of housing, it would be better to predict the probability of housing insecurities. If there were some long texts from the survey or service call/messages, it would be more helpful to analyze the status of the member. For example, after the member interaction, we could get some summarization of interaction results. For these long texts, we could use some SOTA NLP models such as BERT or GPT-3 to enhance our predictions.

7.5. Cost efficiency analysis

We also conducted a brief cost efficiency analysis to ensure that our solution is cost-efficient. Below table is overall cost analysis for each suggested solution. Although this calculation is based on rough assumptions without considering other complex factors, it ensures that our solutions have potential to give benefits to Humana and its members. The following chart briefly explains the expected savings based on our recommendations. In calculating the amount below, we assume that all potential members in housing securities remain in the premium if all the recommended solutions are provided to them, otherwise leave the policy.

	Expected Income	Expected Cost	Expected Profit	Source
Member count	22,300,000			Humana annual report 2021
Housing inscure ratio	4%			Target distribution from data
Target member	892,000			member count * housing insecure ratio
Total premiums and services revenue	73,843,000,000			Humana annual report 2021
Annual average membership cost	3,311			Total revenue / member count
Annual average hospital reimbursement cost	6,898			Humana annual report 2021
Expected income	9,106,736,000			(Annual membership + hospital reimbursement) * members
Direct housing support (Total investment on housing insecurites)		50,000,000		Mentioned on Humana news[8]
Free health check-up for first-time patients		133,800,000		$150(suggested amount) * members
Health Maintenance / Improvement Incentive Program		535,200,000		$50(suggested membership discount) * 12 * members
Psychological Counseling		2,140,800,000		$200(national average cost per session) * 12 * members
Regional environment segment		2,011,460,000		$2255(national annual average ventilation cost) * members
Total	**$ 9,106,736,000**	**$4,871,260,000**	**$ 4,235,476,000**	

[Table 9: Cost Efficiency Summary]

8. Conclusion

With the housing prices increasing, the housing insecurity issue is becoming a primary concern to healthcare service providers. With recent inflation and upcoming economic recession, the risk of experiencing housing insecurities is growing[14][15]. Therefore, it is beneficial for Humana to predict which members are the most likely to face these challenges in the future in order to come up with targeted solutions at an early stage.

Based on the dataset Humana has provided, we have successfully analyzed the dataset, preprocessed the data with feature engineering, and built a predictive model using the LightGBM algorithm. While our model shows high performance in terms of ROC-AUC score, further improvements can be made to enhance the model's prediction performance if the quality of the dataset can be improved in the direction as we mentioned in our recommendation section.

Also, based on the post data analysis after modeling, we suggested some practical solutions including direct financial support and indirect health related solutions based on the insights that we've found. Connecting members with government solutions or local communities and giving reimbursements to members to keep their environments healthy would also be an alternative solution. We highly recommend taking our proposed solutions into consideration when coming up with targeted solutions for the different segments of members.

Meanwhile, Humana is already moving forward to solve the housing insecurity problem. By combining Humana's movement with our solution, Humana will successfully help members with housing insecurities.

9. Reference

[1] *Measuring Housing Insecurity in the American Housing Survey | HUD USER. (n.d.).* https://www.huduser.gov/portal/pdredge/pdr-edge-frm-asst-sec-111918.html

[2] *Housing Instability - Healthy People 2030 | health.gov. (n.d.).* https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries/housing-instability

[3] *The State of the Nation's Housing 2022. (n.d.). Joint Center for Housing Studies.* https://www.jchs.harvard.edu/state-nations-housing-2022

[4] *Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

[5] *lightgbm.LGBMClassifier — LightGBM 3.3.2.99 documentation*. (n.d.). https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html

[6] *GitHub - fmfn/BayesianOptimization: A Python implementation of global optimization with gaussian processes.* (n.d.) https://github.com/fmfn/BayesianOptimization

[7] *sklearn.model_selection.KFold. (n.d.). Scikit-learn.* https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[8] *Humana Expands National Commitment to Affordable Housing With Additional $25 Million Investment. (n.d.)* https://press.humana.com/news/news-details/2022/Humana-Expands-National-Commitment-to-Affordable-Housing-With-Additional-25-Million-Investment/

[9] *Insurer investment in housing reduces healthcare costs, AHIP says. (n.d.). Healthcare Finance News* https://www.healthcarefinancenews.com/news/insurer-investment-housing-reduces-healthcare-costs-ahip-says

[10] *NCBI - WWW Error Blocked Diagnostic*. (n.d.-c). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7525583/pdf/S2056469420000492a.pdf

[11] *Quality of Housing - Healthy People 2030 | health.gov*. (n.d.). https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries/quality-housing

[12] *Housing Issue Brief* https://populationhealth.humana.com/wp-content/uploads/2020/06/Humana_HousingBrief_Final_External_version_2020.pdf

[13] *How Affordable Housing Can Address Harm Caused by Air Pollution. (2022, June 2). Enterprise Community Partners.* https://www.enterprisecommunity.org/blog/how-affordable-housing-can-address-harm-caused-air-pollution

[14] *Improving Measures of Housing Insecurity: A Path Forward* https://www.urban.org/sites/default/files/publication/101608/improving_measures_of_housing_insecurity.pdf

[15] *2022 State of the Nation's Housing report: 4 key takeaways for 2022. (n.d.). Cost of Home.* https://www.habitat.org/costofhome/2022-state-nations-housing-report-lack-affordable-housing

[16] *Wimalasena NN, Chang-Richards A, Wang KI, Dirks KN. Housing Risk Factors Associated with Respiratory Disease: A Systematic Review. Int J Environ Res Public Health. 2021 Mar 10;18(6):2815. doi: 10.3390/ijerph18062815. PMID: 33802036; PMCID: PMC7998657*

[17] *U.S. Concentration of Dentists - Maptitude Infographic*. (n.d.). https://www.caliper.com/featured-maps/maptitude-dentists-map.html

[18] *Center on Budget and Policy Priorities* (n.d.). https://www.cbpp.org/research/housing/research-shows-rental-assistance-reduces-hardship-and-provides-platform-to-expand

[19] *Humana annual report 2021* https://humana.gcs-web.com/static-files/78c99040-2eed-4231-89f9-e12b3e9ec333

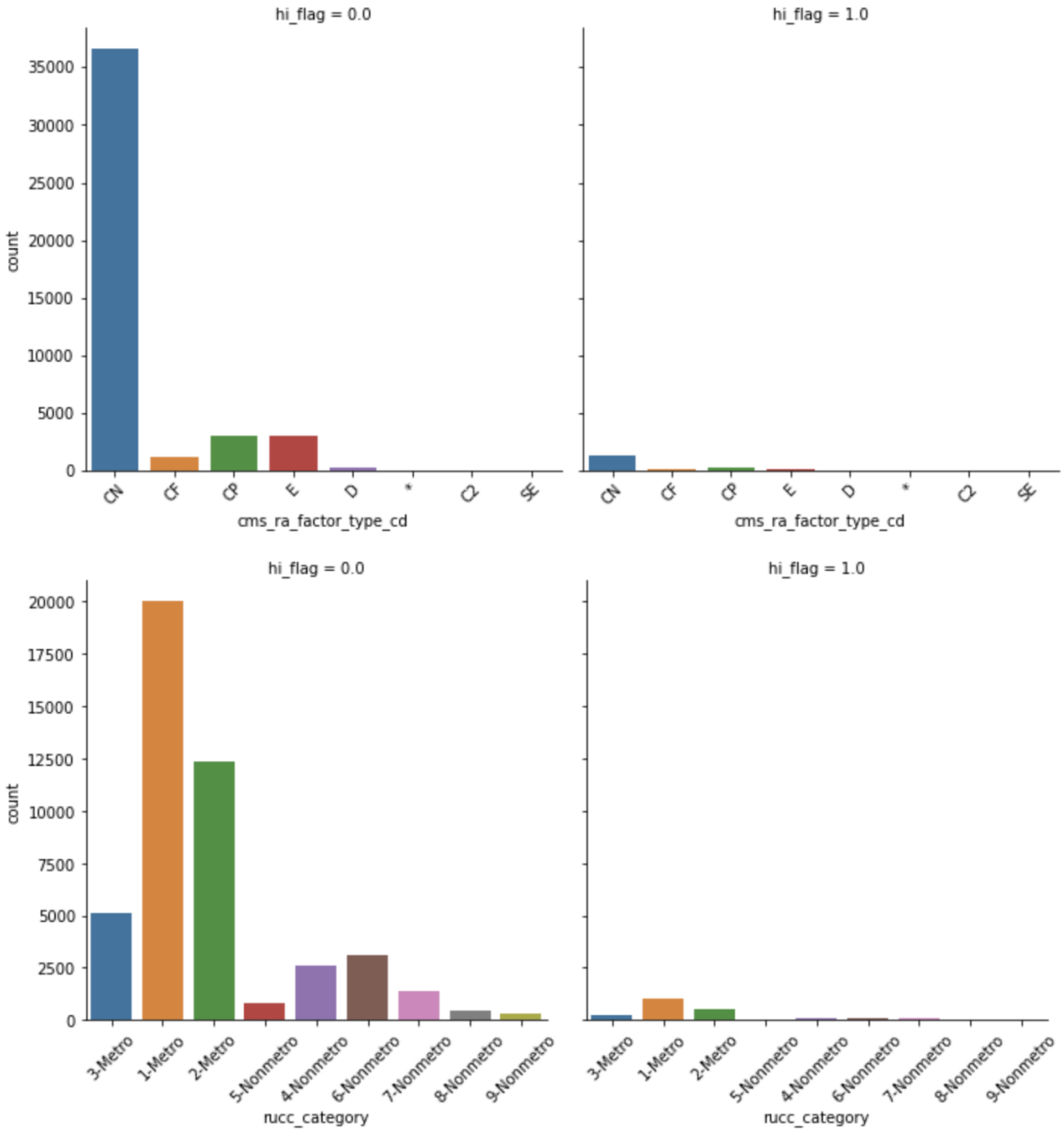
10. Appendix

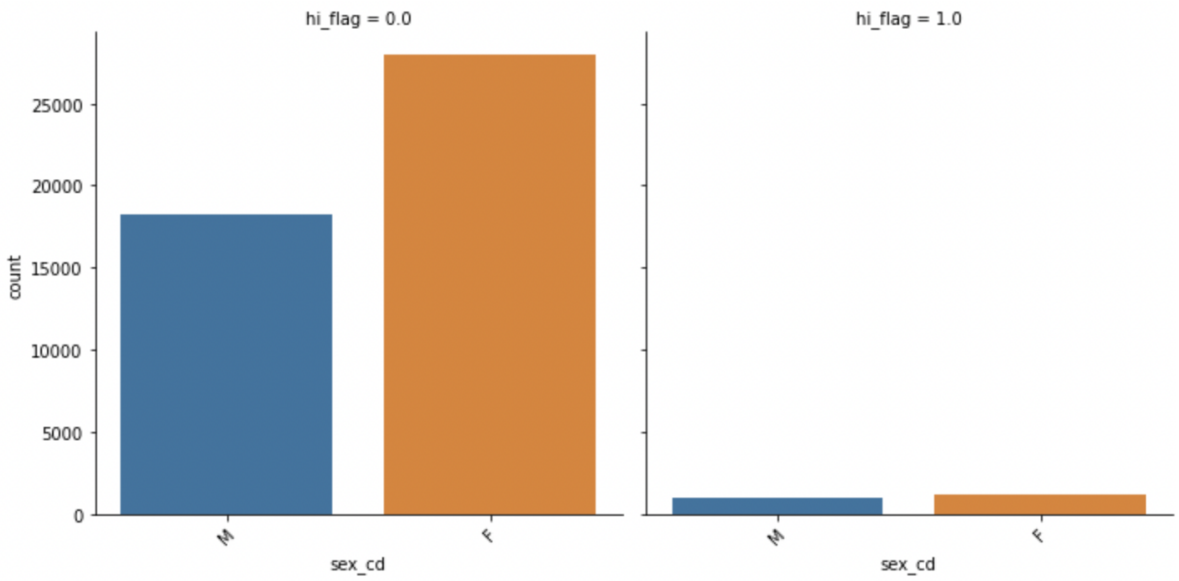
Appendix A. Missing values by columns

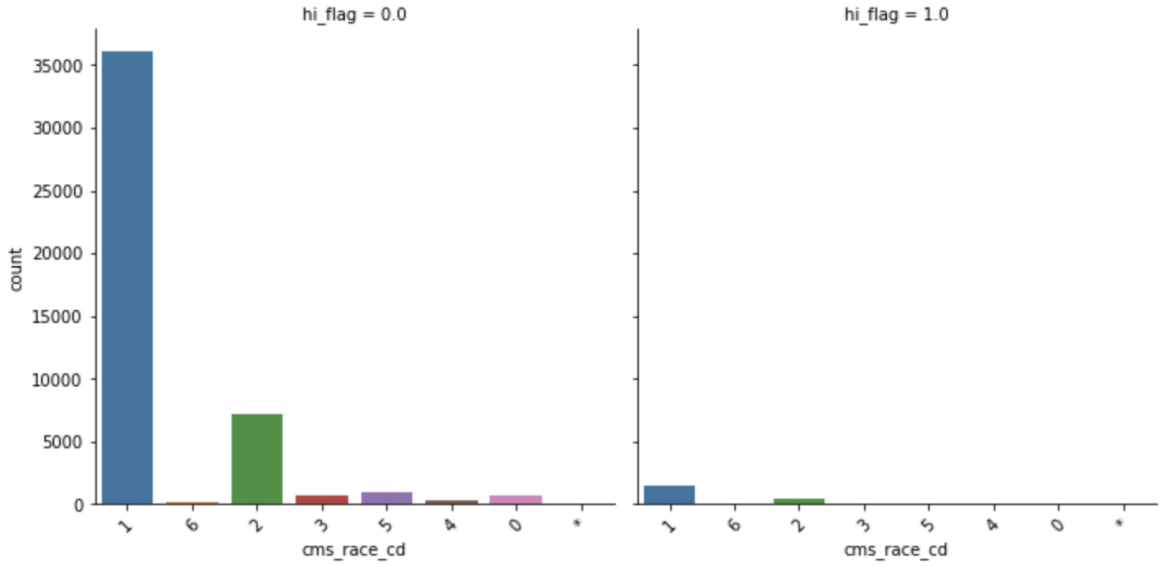
There are 267 columns with missing values

	Missing count	% Missing value
cms_risk_adj_payment_rate_b_amt	60494	0.999570
cms_tot_partd_payment_amt	57585	0.951504
credit_hh_agencyfirstmtg_new	56754	0.937773
credit_bal_autobank_new	56691	0.936732
credit_num_autobank_new	56678	0.936517
credit_num_nonmtgcredit_60dpd	56675	0.936467
credit_prcnt_mtgcredit	56664	0.936286
credit_bal_bankcard_severederog	56646	0.935988
credit_bal_consumerfinance_new	56617	0.935509
credit_hh_autobank	56607	0.935344
credit_bal_nonmtgcredit_60dpd	56592	0.935096
credit_num_1stmtg_collections	54460	0.899868
cons_lwcm10	26739	0.441821
cms_rx_risk_score_nbr	25450	0.420522
cms_ma_risk_score_nbr	24315	0.401768
cms_partd_ra_factor_amt	23813	0.393473
lang_spoken_cd	23785	0.393011
cms_risk_adjustment_factor_a_amt	23641	0.390631
rwjf_homicides_rate	15360	0.253800
rwjf_violent_crime_rate	14832	0.245076
rwjf_child_mortality	14717	0.243176
atlas_snapspth16	14643	0.241953
rwjf_pcp_rate	14626	0.241672
rwjf_men_hlth_prov_ratio	14614	0.241474
rwjf_income_inequ_ratio	14609	0.241391
rwjf_poor_phy_hlth_days	14608	0.241375
rwjf_premature_mortality	14602	0.241276
rwjf_mv_deaths_rate	14580	0.240912
rwjf_teen_births_rate	14576	0.240846
rwjf_premature_death_rate	14567	0.240697
rwjf_poor_men_hlth_days	14563	0.240631
atlas_naturalchangerate1016	14561	0.240598
rwjf_dentists_ratio	14536	0.240185
rwjf_population	14528	0.240053
rwjf_median_house_income	14527	0.240036
atlas_net_international_migration_rate	13949	0.230486
cons_hxmh	13904	0.229742
cons_ccip	13902	0.229709
cons_hxmioc	13878	0.229313
cons_stlnindx	13873	0.229230
cons_hxmboh	13870	0.229180
cons_homstat	13866	0.229114
cons_stlindex	13864	0.229081
cons_mobplus	13854	0.228916
rwjf_std_infect_rate	13645	0.225463
hi_flag	12220	0.201917
rwjf_preventable_ip_rate	12013	0.198496
rwjf_hiv_rate	11087	0.183196
cms_ra_factor_type_cd	3097	0.051173
cms_orig_reas_entitle_cd	2356	0.038929
rwjf_drinkwater_violate_ind	1499	0.024769
atlas_orchard_farms12	1111	0.018358
atlas_pct_diabetes_adults13	489	0.008080
rwjf_food_env_inx	466	0.007700
rwjf_air_pollute_density	410	0.006775
cmsd2_res_res_postop_pmpm_ct	129	0.002132
rx_hum_61_pmpm_ct	108	0.001785
cmsd2_ano_mus_pmpm_ct	107	0.001768
cmsd2_can_mal_end_pmpm_ct	104	0.001718
cmsd2_inf_herpes_pmpm_ct	93	0.001537
cci_hiv_n_pmpm_ct	90	0.001487

Appendix B. Hi-flag distribution by different categorical columns




hi_flag = 0.0
hi_flag = 1.0
count
25000
20000
15000
10000
5000
0
M
F
sex_cd


hi_flag = 0.0
hi_flag = 1.0
count
35000
30000
25000
20000
15000
10000
5000
0
1
6
2
3
5
4
0
*
cms_race_cd