HUMANA-MAYS HEALTHCARE ANALYTICS 2021 CASE COMPETITION

PREDICTION OF VACCINATION HESITANCY FOR COVID-19

1. EXECUTIVE SUMMARY

1.1. STUDY PROPOSAL

This study focuses on helping Humana identify Medicare members who are hesitant to get the COVID-19 vaccine. Our goal was to predict the possibility of vaccine hesitancy for each member and based on the prediction models, we would like to suggest recommendations for Humana to increase access to these members to ensure vulnerable populations receive equitable care.

1.2. MODELING

Before rushing into modeling directly, we spent a considerable amount of time in research trying to understand our business problems and all the features included in the model. Then with a relatively good understanding of all the features as well as our prediction target - Covid-19 vaccination with exploratory data analysis, we picked a series of models that we think would be suitable for this particular case, including Logistic Regression, Lasso regression, XGBoost, and Random forest.

With our selected models, we performed a first round of modeling without feature engineering and feature selection to develop a better understanding of the important features selected by our models. Then we spent a significant amount of time in feature engineering based on our research results and interviews with industry professionals. Then we trained different models and used logistic regression as our benchmark model. We used the AUC score (Area Under the Receiver Operator Characteristic Curve) as our model selection criteria and chose the XGBoost model which has the best model performance. We then performed hyperparameters tuning as our final modeling step.

Through several cycles in feature engineering, feature selection, and hyperparameters tuning, we arrived with an XGBoost model with an AUC score of 0.678. We then conducted further analysis on our model based on feature importance and derived useful business implications and insights for the covid-19 vaccination situation.

1.3. RECOMMENDATION

We suggested two strategies, reducing physical and psychological barriers, to overcome vaccine hesitancy and reach out to the most vulnerable populations. To lift physical barriers, we proposed vaccines by home care and telehealth by Humana’s app. To defeat psychological barriers, we recommended engaging the community by Go365 and associating with healthcare providers.

2. CASE BACKGROUND - THE BUSINESS PROBLEM

Since the first-ever reported case of COVID-19 in November 2019, the Coronavirus has been declared a global health pandemic and a major health emergency all over the world. More than 44.2 million reported infections with COVID-19 and nearly 710 thousand reported deaths from COVID-19 in the U.S. so far. According to a study supported by NHI[1], COVID-19 vaccinations prevented about 140,000 deaths in the U.S. by May 2021, during the first five months they were available. The findings highlight the vital role of COVID-19 vaccines in saving lives and controlling the pandemic. However, vaccines continue to face skepticism. Concerns over their safety and efficacy, often due to distrust and misconception, lead individuals to delay or refuse vaccines. WHO recently named vaccine hesitancy as one of the top threats to global health.[2] As of October 2021, 22% of adults over 18 in the U.S. remain unvaccinated for COVID-19. As a result of lagging vaccinations and the delta variant, COVID-19 cases, hospitalizations, and deaths are on the rise again.

As a top leading health care company, Humana is dedicated to overcoming COVID-19 nationwide by removing financial barriers and ensuring continuous access, especially for the most vulnerable population. Humana announced to cover COVID-19 testing and treatment costs and offered telehealth alternatives for care due to physical distancing requirements. However, increasing treatment costs due to the ongoing pandemic can be a significant burden to Humana. The Centers for Medicare and Medicaid Services (CMS) reports that Medicare fee-for-service COVID-19 hospitalizations average $24,033[3]. According to CDC Morbidity and Mortality Weekly Report, higher hospitalization and death rates were observed among people not fully vaccinated[4]. 92% of the hospitalizations and 91% of the deaths were among persons not fully vaccinated during April 4–July 17, 2021.

As a result, overcoming vaccine hesitancy is beneficial not only for members but also for Humana. Members can decrease their hospitalization-related costs and avoid the risk of severe disease and death. In addition, Humana can utilize saved reimbursements to commit to helping people achieve their best health, which aligns with Humana's mission. These actions can improve customer satisfaction, prevent member churn, and increase the corporate image of Humana. More importantly, since Medicare

[1] National Institutes of Health https://www.nih.gov/news-events/nih-research-matters/vaccines-prevented-140000-covid-19-deaths-us

[2] Clarissa Simas, and Heidi J. Larson, Overcoming vaccine hesitancy in low-income and middle-income regions https://pubmed.ncbi.nlm.nih.gov/34112811/

[3] https://www.cms.gov/files/document/medicare-covid-19-data-snapshot-services-through-2021-03-20.pdf

[4] https://www.cdc.gov/mmwr/volumes/70/wr/mm7037e1.htm

Advantage is mainly for underserved populations, playing a critical role in addressing healthcare inequity can boost Medicare's reputation.

3. TECHNICAL ANALYSIS

3.1. MODEL OBJECTIVE

This year's competition has two clear objectives: to predict the rate of Covid vaccine hesitancy and to reduce the potential bias in the machine learning model. The model will be judged based on important metrics such as AUC-ROC (area under the curve for the receiver operator characteristic curve) and disparity score.

For further explanation about the metrics used in the competition: AUC-ROC is a performance measurement for machine learning problems at various threshold settings, this is measured by true-positive rate and false-positive rate. AUC-ROC is a frequent metric to compare the performance of machine learning models with each other, with the higher the AUC score is, the better the model will do at classifying the positive class in the dataset. For this competition, the scoring metric will predict a positive rate and the outcome will be the sum of true positives and false positives.

The disparity score is an important factor in measuring the bias in a machine learning model. Humana has chosen a formula to calculate the disparity score ratio (DR) as:

DR = Sn / So

Where Sn is the scoring metric for each class and S0 is the scoring metric for the reference group. The reference group is defined as the privileged group within a sensitive variable class.

Discrimination and bias in machine learning have haven a prevalent problem in building prediction with machine learning. "Technology is a powerful force in our society. Data, software, and communication can be used for bad: to entrench unfair power structures, to undermine human rights, and to protect vested interests. But they can also be used for good: to make underrepresented people's voices heard, to create opportunities for everyone, and to avert disasters."[5] With the tool and the data, we can create predictive analytics systems such that if the history is biased, the system will be codified and create even more feedback loops and echo chambers. We have to be even more careful when implementing prediction models into the health care system, which is intended to have the goal of treating people's health equally so as not to create any dangerous bias to the patients.

[5] Martin Kleppmann, Designing Data - Intensive Application (Company of publication: O'Reilly Media, Inc, 2017), book introduction page.

3.2. THE DATASET

- **Training data**: 974,842 records by 366 variable columns, plus ‘covid_vaccination’ target variable
- **Holdout data**: 525,158 records by 366 variable columns

The dataset consists of Medicare Advantage, Medicare Advantage prescription drug (MAPD) and Medicare Prescription Drug Plan (PDP) members, variable columns representing their social determinants of health, and a binary response variable “covid_vaccination.” This response variable is a binary flag to indicate each member’s vaccination status: A value of “1” represents if the participant has already been vaccinated and “0” represents if the participant hasn’t been vaccinated.

Besides the target variable, the social determinants of health variables included in the dataset range from medical and pharmacy claims information to credit and consumer data. They come in a variety of forms: scalar, integer, binary categorical, and multi-class categorical. The dataset is imbalanced relative to the “covid vaccination” response, with only 17.38% of records in the training data being in the positive class of participants having been vaccinated (1), and 82.62% haven’t been vaccinated (0).

This year we have 2 other variables to calculate the disparity point for bias in machine learning models: SEX_CD and RACE_CD, representing the genders and races of participants. To calculate the disparity score, we have to take the AUC score and for each of the categories in the variables and compare them with the reference value. The reference group for these 2 variables will be “male” when the value of “SEX_CD” is 1, and “white” when the value of “RACE_CD” is 1. The reference group is highly unbalanced in the variables, which will pose some challenges when balancing the bias in the model.

3.3. DATA CLEANING AND IMPUTATION

Due to the untidy dataset with many missing or null values, we split the data into two categories: scalar variables and categorical variables.

Missing values

Overall, Humana’s raw data is relatively complete in terms of missing values: Out of the 366 columns, the average missing data is about 2%, and there are only 64 columns that are missing more than 5% of data. We closely inspected each of the 64 columns that are missing the most data to understand the impact of these features to our target and the reason behind a large number of missing values, then customized ways such as imputing the missing values or dropping the entire column.

For scalar features, missing values and “*” are replaced by the median values for that variable. For categorical features, we impute the null values with a label “Unknown” to maintain the information and minimize the bias.

Taking one variable as an example, the “language spoken” feature is missing more than 70% of data, and due to its categorical nature, imputing the missing value could potentially introduce a lot of bias to our model so we decided to drop the feature. We also noticed that there are more than 10 features related to the census data that are missing the same percentage of data, which could be due to issues in the data collection method or differences in sample demographics, therefore we took a safer approach to encode the missing values with label “Unknown” to minimize the bias we introduced.

Data type inconsistency

A lot of features have the problem of data type inconsistency, and it is crucial to deal with this problem before training our model. We used a function to detect inconsistent data types, and the majority of this problem comes from numerical data in a string format and missing values represented by “*”. For example, a feature called (rx_gpi4_6110_pmpm_ct) contains ‘0.66667’ as a string and “*” as null value, and the meaning of this feature is the count per month of prescriptions related to AMPHETAMINES drugs in the past year. For features like this that should only have numerical data, we replaced the missing values and converted all data types to float.

One-hot encoding

One-hot encoding is used to convert categorical string data into numerical data with 0 and 1. For each category in a feature, we created a separate column with 1 representing the category and 0 not representing the category. As for the missing values, to not introduce too much bias into our model, we created a new category called “Unknown” for each categorical feature.

We realized that there are many features with categories only representing a very small percentage of the data, for example, 0.1%. In order to keep our dataset tidy and our model interpretable, we used a function to drop all the categories that represent less than 5% of the sample data. This technique also took care of the collinearity problem, since for each feature, we can treat those dropped categories as one category as “Others”, which carries the meaning of minority categories who represent a very small portion of the sample data. Dropping this combined category for each feature simulates the process of “drop_first = True” in one-hot encoding.

3.4. RESEARCH AND FEATURE ENGINEERING

In our feature engineering part, we realized that with this vast amount of features (600+ features after one-hot encoding), a lot of noise would also be included in the model. Therefore, we decided to first subset the most important features using XGBoost’s feature importance, and use those features in feature engineering to generate meaningful and potentially useful new features. More research results will also be covered later in the business analysis and implication part.

After a cross-comparison between XGBoost’s feature importance and Pearson correlation significance, we subsetted a group of 77 most important features. Then we conducted a vast amount of primary and secondary research to understand the features and to find the underlying groups. Based on our research and feature importance from our XGBoost Model, we divided the most important features into three main categories: population demographics, resource accessibility, and pharmacy claims.

Two major feature engineering techniques were applied to the model: feature transformation and feature combination. Feature transformation works by studying the feature itself and creating new meaningful features from this one feature such as reciprocal, square, square root, and logarithm. Feature combination works by studying different features, exploring the interaction between them, and then generating new features from them.

Population Demographics

Based on our primary research and feature importance, population demographics information contains some of the most important factors that influence people’s decision on whether to get vaccinated. This also makes sense intuitively: demographic factors such as age, education level, income, location, even political orientation can affect people’s decision on whether or not to get vaccinated.

A lot of features in this group can be generalized in feature combination with features from other groups. For instance, features like ‘Per capita income’, ‘Household food insecurity’, and ‘Census Geo-unit quality score’ were used as denominators or multipliers in the feature combination steps with other features.

Resource Accessibility/Availability

The level of resource accessibility & availability is another important section that can have a profound impact on people’s decision on vaccination. For example, whether a person owns a car, whether he/she is covered by insurance, whether there's a vaccination location available nearby, or whether they're farmer’s markets that provide healthy food and vegetables.

A lot of new categorical features were created using interactions between original features. For example, we can build a new feature combining whether or not a person is a mail buyer (binary categorical feature) and the likelihood of a person will self-monitor illness (categorical feature), we can engineer new categorical features and form hypotheses around those new features. A person with access to mail orders (cons_mobplus_M == 1) and has a score of 9 in illness self-monitor (cons_hxmioc == 9) is most likely to be able to take care of themselves and receive proper treatment and medicine.

Pharmaceutical Claims

Among the feature importance list generated from XGBoost, more than 30 features are related to the pharmaceutical claims category. Examples include cost per month on prescribed drugs (numerical feature), hospital visits per month on non-behavioral per month (categorical feature). We generated a lot of interaction features based on this category. This technique is very useful to subset different groups of people. For example, people with a high number of hospital visits per month who’s also among the top 25 percentile in terms of cost on prescribed antihyperlipidemic drugs per month can be grouped into a new categorical feature, representing a group of people who have low-density lipid (LDL) cholesterol issues and are frequently making hospital visits.

Auto Feature Augmentation (not included)

After manually generating features we think would be meaningful, we wanted to include as many new features as possible that would potentially be helpful in our model. We used packages called AutoFeat and FeatureTools that automatically generate new features based on existing features using similar ideas in feature transformation and feature combination. However, a problem associated with these automation tools is that they can not be customized for each feature, thus creating a lot of reductant and meaningless features. Besides, the number of new features grows exponentially when we introduce the second layer of feature combination which allows the package to generate features based on the interaction between features. Theoretically, a two-layer auto-feature function with 77 original features would generate around 81,655 features, each with the original 974,842 samples. Therefore we deemed this approach to be invalid and adhered to our manually engineered new features.

In summary, based on our study for the three major categories above as well as our research results, we generated 2500+ new features from feature transformation and feature combination techniques.

3.5. MODEL SELECTION

Due to the classification nature of the target variable (Covid_vaccination), a vast amount of data around 1 million samples with 3000 features each, and the non-linear high dimensional interactive relationship between the features and the target, traditional statistical models such as Logistic regression and Lasso regression may not be our best suit. Rather, tree-based models, with their nature of grouping features and samples together, and a high tolerance for non-linear relationships, would come in as an ideal fit for our classification model.

In the tree-based models, decision tree is the most simple model, however, the performance may suffer from various factors including bias, sample size, and high variance. Bagged trees build several decision trees together on bootstrapped training samples, which could improve the model performance. Random forest, on the other hand, is superior to bagged trees by using a technique to decorrelate the trees. In random forest models, each time a split in a tree is considered, a randomly chosen group of predictors will be chosen as split candidates from the full set of predictors. This allows random forest models to decorrelate between the trees and produce a lower variance when we average the trees. Boosted trees work similarly to bagged trees, except that each tree is grown in sequential order: each tree is grown using information from the previous trees, which comes in handy in our cases as it could be more efficient compared to random forest models.

Our model selection result is also in line with our empirical experiment results. Out of all the models we tested, XGBoost has the best performance of 0.678 in terms of AUC score; Random forest returned a score of 0.664; Lasso regression returned a score of 0.642, and Logistic regression with a score of 0.651.

With all factors considered and numerous training session results, we decided to use the XGBoost package because of its superior performance and relatively high efficiency. XGBoost provides a regularized gradient boosting framework with the following features that particularly suits our case situation: it has clever penalization of trees, extra randomization parameter to minimize variance, and automatic feature selection.

3.6. FINAL MODEL CONSTRUCTION

With our new engineered features along with the original features, we performed hyperparameters tuning for our final model. We performed grid search cross-validation to find the best set of hyperparameters. While this technique would try each possible combination of hyperparameters, it would be extremely inefficient and time-consuming. Therefore, a technique called early stop is incorporated to improve training efficiency. The rationale behind the early stop is: after a certain number of iterations set by the user (we used 10), if there's no improvement on model performance, the grid search will skip the combinations of hyperparameters that deems to have worse performance than the existing set of hyperparameters. Using grid search and the early stopping technique, we search for the listed hyperparameters below:

- **Max_depth**: the maximum depth of each tree in the model. This is usually one of the most common and important hyperparameters in tree-based models as high values are often associated with overfitting, and it's also a measure of model complexity.
- **Gamma**: the minimum error or loss reduction for a leaf node to develop into new branches. The larger gamma is, the more conservative the algorithm will be, and the fewer depth each tree would develop.
- **Reg_alpha**: an L1 regularization term on weights (similar to Lasso regression). Can be used to reduce high dimensional data and as a form of feature selection in the model. A high alpha would result in a more conservative model.
- **Reg_lambda**: L2 regularization term on weights (similar to Ridge regression). Similar to alpha, lambda can be used to regularize the model.
- **Colsample_bytree**: a subsampling hyperparameter that occurs whenever a tree is constructed, which takes the subsample ratio of columns.
- **Min_child_weight**: the minimum sum of weights of all observations required in a child. It's a threshold to prune branches from a large tree. This is a regularization hyperparameter used to control overfitting and generalize the model.
- **N_estimators**: the number of trees in each model. A large number of trees for the model could lead to a longer training time but lower variance.

With our best set of hyperparameters from grid search cross-validation, our final model returned 255 features with non-zero feature importance out of the 3000+ features we put into the model. As XGBoost already included a feature selection step in the algorithm and considered our study objective to generate interpretable results to help with business strategies, we did not perform dimension reduction techniques like principal component analysis and nonnegative matrix factorization. Model interpretation and business implication were answered by feature importance and SHAP analysis which were covered in the next section of our report.

3.7. MODEL EVALUATION

We produced an XGBoost model with an AUC score of 0.678, while in the cases of classification models, it is very important to analyze other model performance metrics to make sure the constructed model can deliver desired results. We used a confusion matrix to understand our model performance.

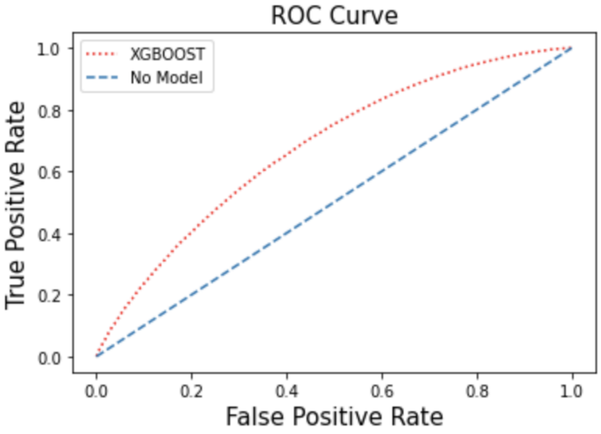
Sensitivity: sensitivity is a measure of a model's ability to detect true positive cases. It is calculated by dividing true positives by total positives. Our model has a sensitivity score of 66%, which means our model can identify about 2/3 of people who have not been vaccinated as a non-vaccinated population.

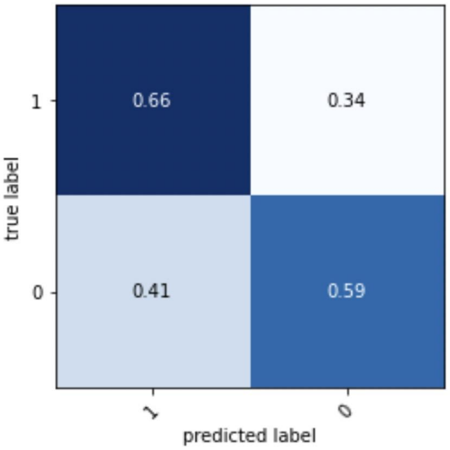
Specificity: specificity is a measure of a model's ability to detect true negative cases. It is calculated by dividing the true negatives by total negatives. Our model has a specificity score of 59%, which means our model can identify about 59% of people who have been vaccinated as a vaccinated population.

Accuracy: Our model accuracy of 86.6% is not particularly high, as we are using an unbalanced dataset with about 17% positive sample target.

Both our specificity and sensitivity do not have outstanding performance, but none of them has extremely low performances, which suggests that our model is not particularly biased towards one direction. Given our business case background, our objective would be to over predict positive cases (people who have not been vaccinated) to develop insights for the reason of vaccination hesitancy and create business strategies accordingly. Therefore, our 66% sensitivity score suggests that our model is mediocre and has room for further improvement.

```
No vaccination: ROC AUC=0.678
```





4. KEY PERFORMANCE INDICATOR ANALYSIS

- **XG boost Feature Importance**

We used XG boost feature importance function called get_booster() to get the most important features after tuning and training the model. We used the calculated numerical value of "gain" to take each feature's contribution to each tree in the model. The formula to calculate "gain" can be shown as:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$ [6]

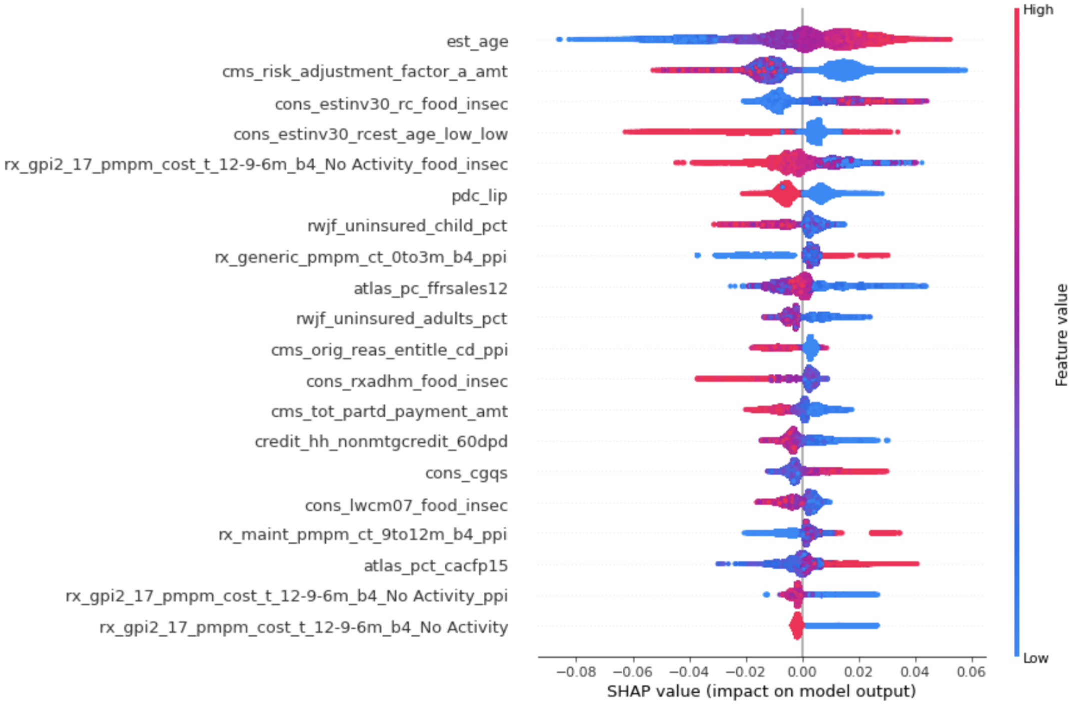
The model can be decomposed as: comparing between the scores of the new left leaf, the new right leaf, the additional new leaf and adding in the regularization. The formula has the same basis as the "pruning" technique and will search for optimal splits in the model.

- **SHAP Summary plot**

In our feature importance analysis we have incorporated XGBoost feature importance analysis and SHAP plot (Shapley Additive Explanations) which incorporates well with tree-based algorithms. It's important to note that we have included other feature importance tools in the initial exploratory data analysis and we can compare between the features that repeatedly appear in the feature importance ranking. SHAP values is a well-known method in post-model analysis to compare and analyze the final

[6] From XGBoost documentation: Introduction to boosted tree - https://xgboost.readthedocs.io/en/latest/tutorials/model.html

features, since it generates numeric values (SHAP values) which can be used to calculate the important role of the features to the model. Based on these SHAP features, we can identify 20 key factors, with demographic, pharmaceutical, and credit data playing the most essential roles in the model.



In our SHAP features we are integrating important variables made up of both the original variables from Humana dataset and our synthetically created features. We will choose 8 most important variables and make explanations incorporated with our research:

1. est_age: This is the variable that appears multiple times between feature importance sessions by different models. We can see that age has a positive correlation with Covid vaccination rate, which means that the lower the age of a participant, the lower the possibility that they will have Covid vaccination.

2. cms_risk_adjustment_factor_a_amt: Medicare Risk Adjustment Factor (RAF) pay plans for the risk of the beneficiaries they enroll, instead of calculating an average amount of Medicare/Medicare Advantage beneficiaries, with low value

indicating a lower health-related risk. We can understand that if a patient already has high health risks, they will be more reluctant to get Covid vaccinated.

3. rx_generic_pmpm_cost_6to9m_b4_food_insec: This is our newly created variable by comparing cost per month of prescriptions related to generic drugs in the past sixth to ninth month prior to the score date to Household food insecurity (%, three-year average), 2013-15.
4. cons_estinv30_rcest_age_low_low: This is our newly created variable, which represents the lower 25th percentile of participants' age with lower 25th percentile of estimated household investable assets - in other words, the population younger than 67 with low household assets will have a lower chance to get the vaccination from the SHAP values.

5. rx_gpi2_17_pmpm_cost_t_12-9-6m_b4_No Activity_food_insec: The trend of cost per month of prescriptions related to VACCINES drugs in the past sixth to the ninth month versus ninth to the twelfth month prior to the score date in relation with the percentage of households with very low food security in the area between 2013-2015. According to the 2 vaccine hesitancy research: Vaccine hesitancy: the next challenge in the fight against COVID-19[7] and COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates[8], whether or not a participant has been influenza vaccinated will have a huge impact on Covid vaccination hesitancy, and it explains why this variable is in our top 5 in feature importance.

6. pdc_lip: proportion of days covered for prescriptions related to hyperlipidemia in the past one year. The SHAP graph means that the more prescription days that a patient has, the more probable that they won't get vaccinated.

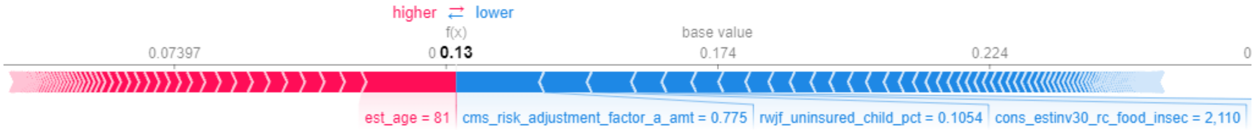
7. rwjf_uninsured__child_pct and Rwjf_uninsured_adults_pct: Clinical Care - taken by geography level and sourced from Robert Wood Johnson Foundation data, meaning that the participant is living within an area with a high percentage of population under age 65 without health insurance. By incorporating the 2 variables we can see that they have negative correlation with Covid-19 vaccination rate.

[7] Vaccine hesitancy: the next challenge in the fight against COVID-19, https://link.springer.com/article/10.1007/s10654-020-00671-y

[8] COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7920465/

8. rx_generic_pmpm_ct_0to3m_b4: count per month of prescriptions related to generic drugs in the past three months prior to score date. This variable positively correlates to the possibility of Covid vaccination, with the more someone spends on generic drugs the more change that they will get vaccinated.

Diving deeper into SHAP analysis, we can plot an observation and understand the predictors that drive the result from our model. This is one observation that has the same prediction as our test data, and we can see the valuable features that drive the accuracy:



- The output value is the prediction Covid vaccination probability (the prediction of the first row is 0.13).
- The base value is the prediction for Covid vaccination probability given that there are no features (which essentially means the average output of test data).
- Red/blue: Features that push the prediction higher (to the right) are shown in red, and those pushing the prediction lower are in blue.
- cms_risk_adjustment_factor_a_amt will have the most positive impact on the Covid vaccination rate, for this particular observation it has the value of 0.775.
- est_age has a negative correlation with the Covid vaccination rate, in this observation, the patient has the age of 81 years old.

5. BUSINESS ANALYSIS AND RECOMMENDATIONS

5.1. INTRODUCTION

COVID-19 pandemic showed that the failure to protect the most vulnerable members of society not only harms them but also increases the risk of spread of the virus, with devastating health consequences for all.[9] Humana recognized the importance of health equity in advance and took various actions to outreach vulnerable populations. Humana announced to cover COVID-19 testing and treatment costs and offered telehealth alternatives for care. Moreover, Humana established a clinical outreach team to engage with the company's most vulnerable populations and actively

[9] COVID-19 and Health Equity—A New Kind of "Herd Immunity" https://jamanetwork.com/journals/jama/fullarticle/2766096

cooperate with the government in the vaccination process. For their efforts to offer members quality care at the right essential time, Humana led the health insurance industry as they ranked 1st in the 2020 American Customer Survey Index. In this section, we would like to suggest actionable insights and recommendations for Humana to create greater value.

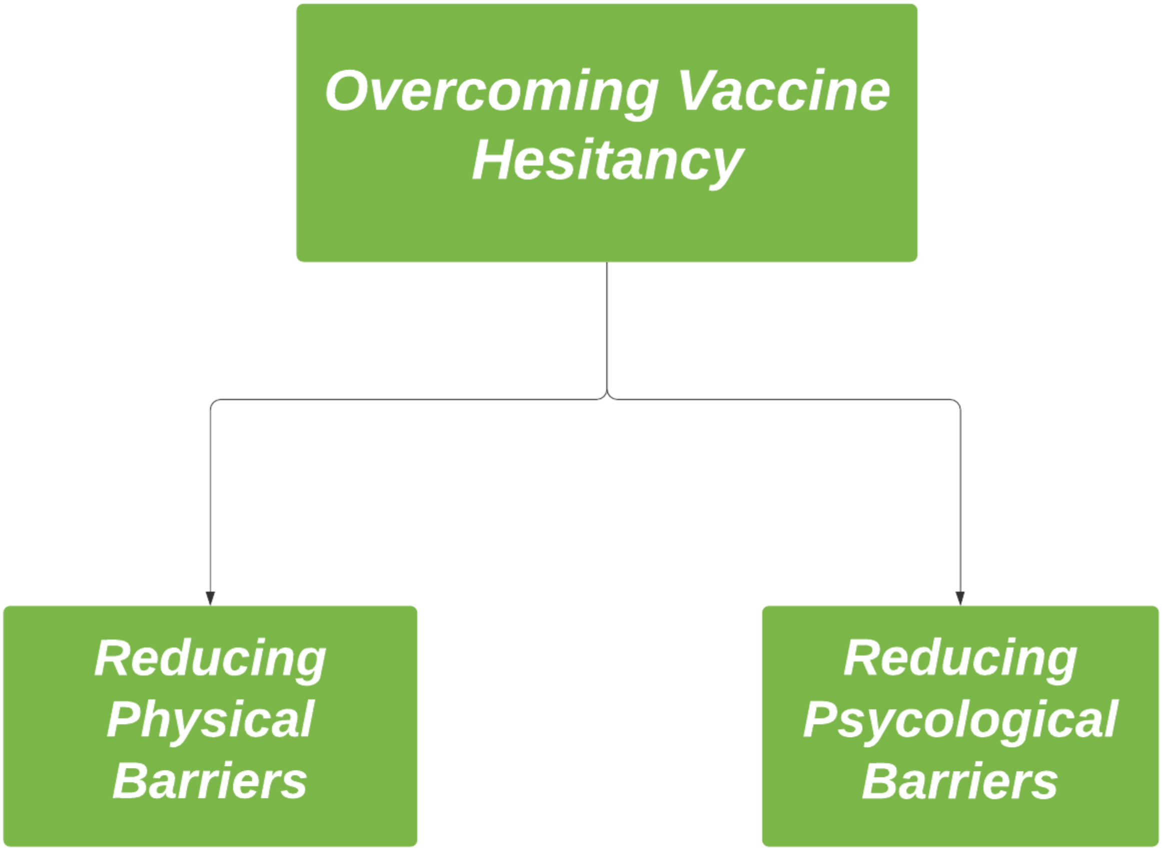
5.2. REASON FOR HESITANCY

It is critical to developing tailored strategies to increase the acceptability of the COVID-19 vaccine and decrease hesitancy. Vaccine hesitancy results from a complex decision-making process, influenced by a wide range of contextual, individual and group, and vaccine-specific factors, including communication, socioeconomic, geographic barriers, experience with vaccination, risk perception, and design of the vaccination program[10]

As seen from our analysis, people with disabilities, low assets, high debt, low food security, and those who suffer from hyperlipidemia (high cholesterol) are mostly reluctant to get vaccinated. Moreover, features interact with each other according to our analysis. For example, the majority of disabled belonged to the low assets group. For these members, there are greater barriers than the others. They might have fewer resources, lower access to vaccine locations, and less information about COVID-19.

We defined the Covid-19 vaccination hesitancy as physical barriers and psychological barriers for these groups. Physical barriers can be explained as lower access to vaccines. Members with disabilities are likely to suffer from a lack of mobility. Psychological barriers can be explained as misconceptions and distrust. Members with low assets, low income, and high debt might live in communities where distrust is widespread, or have lower resources to get accurate information about vaccination.

[10] Factors Associated with COVID-19 Vaccine Hesitancy
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8004673/


Overcoming Vaccine Hesitancy
Reducing Physical Barriers
Reducing Psycological Barriers

5.3. OVERCOMING PHYSICAL BARRIER

Home Care: Vaccine at home

It may be challenging to access vaccination sites for people with disabilities and the elderly due to lack of mobility. According to our analysis, people with disabilities have fewer assets, lower income, and higher debt on average. They also have more chronic diseases and obesity, causing severe illness due to COVID-19. Moreover, they might have difficulty wearing a mask, sanitizing their hands, and keeping social distance. Since these conditions put people with disabilities at higher risk, they are one of the most vulnerable groups to be prioritized, deeply needing help.

Currently, Humana is providing home health care coverage, as part of the government mandatory addition to medicare. Despite the pandemic's initial reduction in demand for home care services, aging, greater illness such as circulatory and heart disease, and strong government support keep the industry growing. Industry revenue has grown at an annualized rate of 3.3% to $109.6 billion over the five years to 2021.[11] In-home care is expected to be a preferred option for members who avoid being at the medical facility.

Our recommendation is to help people with disabilities get the vaccine at home by home care providers. Humana can reach out to those members and arrange appointments with the caregiver. These efforts will potentially increase customer satisfaction and reinforce a positive impression about in-home care.

Telehealth: MyHumana

Telehealth allows hospitals to provide quality healthcare services with improved access at lower costs. COVID-19 pandemic boosted demand for telehealth services since infection limited patients to receive in-person healthcare services. Many physicians started to use telehealth as a substitute that avoids the risk of patients contracting or spreading the virus. Insurers promote telehealth to lower medical expenses. As a result, industry revenue has increased at an annualized rate of 30.4% to $3.2 billion over the five years to 2020, including an increase of 9.7% in 2020 alone.

Medicare Current Beneficiary Survey (MCBS):[12]

- 83% of beneficiaries reported having access to the Internet.
- 70% of the beneficiaries responded that smartphones are the most commonly used type of technology.
- 64% reported that this provider currently offers telephone or video appointments among beneficiaries who have a usual health care provider.

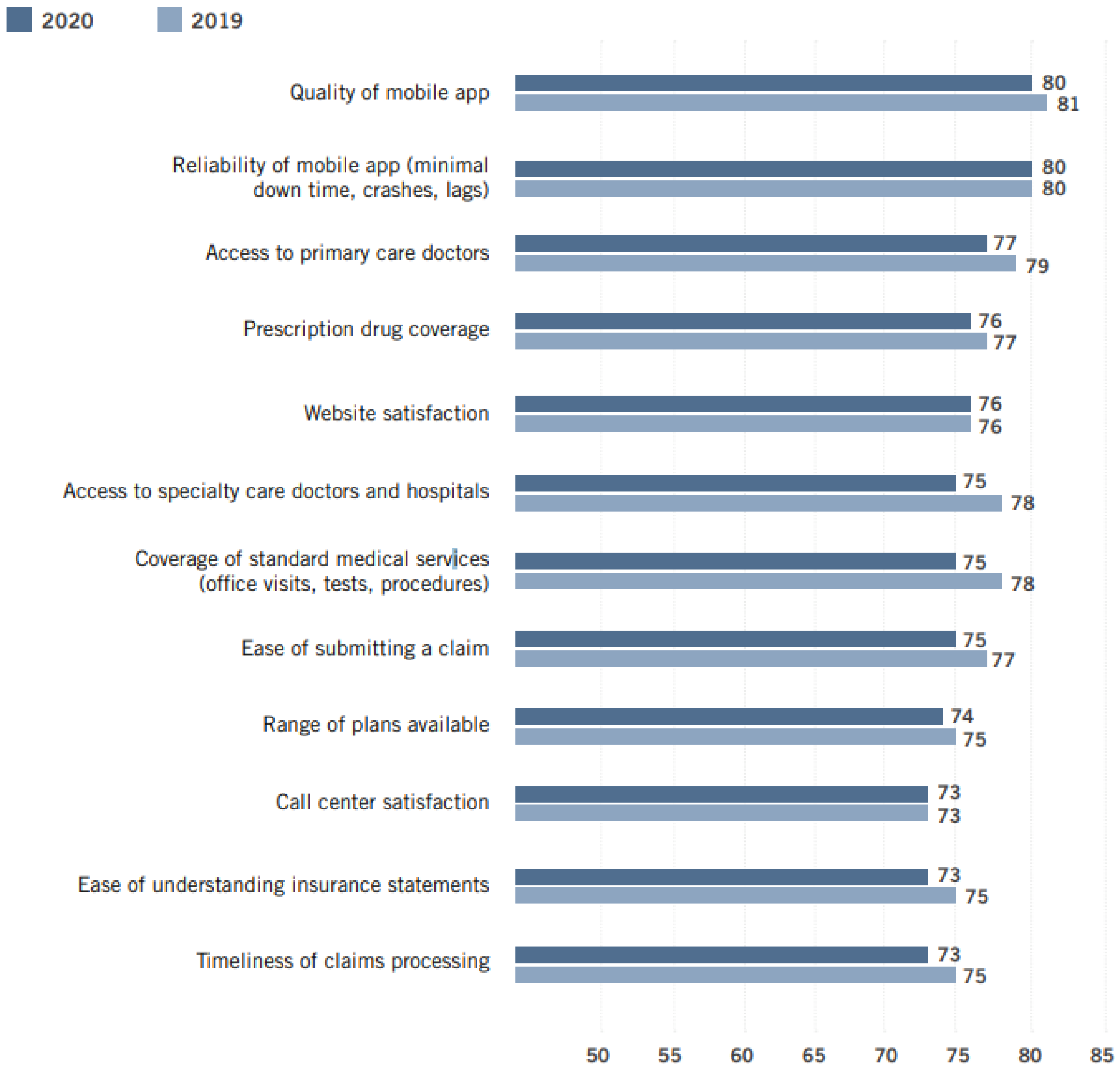
According to the Medicare Current Beneficiary Survey conducted fall 2020, 83% of Medicare members have access to the Internet, and 70% of them use smartphones the most. Therefore, tech devices including smartphones can be used to reduce the distance from vaccines for the members hesitant to get vaccinated.

[11] IBIS https://my-ibisworld-com.ezproxy.bu.edu/us/en/industry/62161/about

[12] Medicare Current Beneficiary Survey (MCBS) 2020

HEALTH INSURANCE

Customer Experience Benchmarks
Year-Over-Year Industry Trends



ACSI (0-100 Scale)

© 2020 American Customer Satisfaction Index. All Rights Reserved.

Our recommendation is to use MyHumana mobile app to outreach reluctant members. By adding the ‘Vaccine’ category at the ‘Provider Finder’, it is possible to push alerts when a vaccine is available nearby. In addition, by building networks with health providers, members might be able to make reservations using the app and check the schedule.

The 2020 ACSI health insurance survey shows the importance of mobile apps as a measurement of customer satisfaction. Mobile apps continued to gain the highest ratings, followed by websites. If members have positive and comfortable experiences while using apps, it will increase not only vaccine acceptance but also customer loyalty.

Moreover, members might be able to check their symptoms through apps after getting vaccinated, and those who suffer from side effects could be followed up by telehealth. Members could alleviate the concern and burden of side effects throughout this process.

5.4. OVERCOMING PSYCHOLOGICAL BARRIER

Engaging Community

On the basis of our data analysis, we observed that vaccination rates differ from region to region. Widespread misconception and distrust in the community are likely to make people reluctant to get vaccinated. Psychological science indicates that people are affected by the behavior of the neighborhood and social norms can influence vaccination acceptance behaviors. In a study focused on flu vaccination among adults, decisions to vaccinate were influenced by their social circles' actions and the consequences of the flu without immunization[13].

Therefore, our suggestion is to engage the community to build trust, and increase collaboration.

- Humana Go365
- Humana Vaccine Representative

Humana Go365 encourages members to engage in community events and wellness activities, building community members' strong relationships. Humana can reach out to those reluctant, vulnerable communities by Go365. Humana can select community leaders as vaccine representatives to deliver messages and empower vaccine recipients to share their reasons for vaccination throughout Go 365 activities. Encouragement throughout these healthy activities will be able to reconstruct members' thoughts about vaccines; rather than focusing on side effects, it can be considered something as healthy and familiar.

[13] Building Vaccine Confidence Through Community Engagement https://www.apa.org/topics/covid-19/equity-resources/building-vaccine-confidence.pdf

Associating with Healthcare Providers

Based on our analysis, people using doctors as a source for medical information have less vaccine hesitancy on average. Building trustworthy, continuous, and long-term relationships between healthcare providers and patients is one of the key factors in terms of vaccine hesitancy and business success. Humana has been attracting members by building networks with a broad, sufficient number of high-quality providers.

Thus, by associating with local hospitals, clinics, and primary doctors, Humana will be able to tackle vaccine hesitancy. Humana educates providers about the vaccine's effectiveness, efficacy, and importance, and then providers encourage their patients to be vaccinated. According to the Medicare Current Beneficiary Survey, the two most common reasons medicare members would not get a COVID-19 vaccine are side effects or safety concerns (41%) and distrust (41%). Trustworthy providers can be an effective method in convincing patients to communicate transparently about the process for authorizing, approving, and monitoring the safety of vaccines.

5.5. EXPECTED VALUE FOR HUMANA

Cost-Benefit Analysis

Index	Subject	Number	
A	Fully Vaccination rate (March 2021)	13.04	
B	Number of Humana Medicare Advantage Members (2020)	4,600,000	
C	Number of Fully Vaccinated Members	600,000	A*B
D	Number of Not Fully Vaccinated Members	4,000,000	(1-A)*B

Benefit (Reimbursement amount saved) : **$1,489,200,000**

E[14]	Average Medicare fee-for-service COVID-19 hospitalizations	$24,000	
F[15]	Medicare Covid Hospitalizations per 100K	1,825	
G[16] (*)	Effectiveness of full vaccination for preventing hospitalization	0.85	
H	Preventable Costs of Unvaccinated COVID-19 Patients	$1,489,200,000	D*E*G*(F/100,000)

*Among adults aged ≥75 years, the effectiveness of full vaccination for preventing hospitalization was 91% for Pfizer-BioNTech, 96% for Moderna, and 85% for Janssen COVID-19 vaccines(CDC).
We will use 85%, the lowest.

Cost (Vaccination) : **$360,000,000**

I(**)	Vaccination Cost per Person without disabilities	$80	
J (***)	In-Home Vaccination Cost per Person with disabilities	$150	
K (****)	Percentage of people with disabilities in the dataset	0.25	
H	Total Cost for Vaccination	$360,000,000	D*(1-K)*I + D*K*J

** $40 Medicare payment rates for single-dose. We will assume 2 doses for every vaccine.

[14] Centers for Medicare and Medicaid Services (CMS) Report https://www.cms.gov/files/document/medicare-covid-19-data-snapshot-services-through-2021-03-20.pdf

[15] https://www.cms.gov/files/document/medicare-covid-19-data-snapshot-services-through-2021-03-20.pdf

[16] https://www.cdc.gov/mmwr/volumes/70/wr/mm7032e3.htm

*** Medicare’s additional payment amount for administering the COVID-19 vaccine in the home for certain Medicare patients is $35 per dose.(by CMS.gov)
**** 'cms_orig_reas_entitle_cd'==1 (we didn’t calculated both for disabled and ESRD)

Total Net savings : $1,129,200,000

6. CONCLUSION

In predicting whether or not a Humana member has received a covid-19 vaccination hence to develop insights about people’s hesitancy to covid-19 vaccination, we narrowed over 3000 features down to 255 features, and chose the XGBoost model with an AUC score of 0.678 as our final model. By Interpreting important features, we identified characteristics of members reluctant to get vaccinated as lower income, higher health risk, disabled. Additionally, we proposed recommendations to reduce physical and psychological barriers.

7. REFERENCE

[1] National Institutes of Health
https://www.nih.gov/news-events/nih-research-matters/vaccines-prevented-140000-covid-19-deaths-us

[2] Clarissa Simas, and Heidi J. Larson, Overcoming vaccine hesitancy in low-income and middle-income regions https://pubmed.ncbi.nlm.nih.gov/34112811/

[3]
https://www.cms.gov/files/document/medicare-covid-19-data-snapshot-services-through-2021-03-20.pdf

[4] https://www.cdc.gov/mmwr/volumes/70/wr/mm7037e1.htm

[5] Martin Kleppmann, Designing Data - Intensive Application (Company of publication: O’Reilly Media, Inc, 2017), book introduction page.

[6] From XGBoost documentation: Introduction to boosted tree -
https://xgboost.readthedocs.io/en/latest/tutorials/model.html

[7] Vaccine hesitancy: the next challenge in the fight against COVID-19

[8] COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates

[9] COVID-19 and Health Equity—A New Kind of “Herd Immunity”

[10] Factors Associated with COVID-19 Vaccine Hesitancy

[11] IBIS https://my-ibisworld-com.ezproxy.bu.edu/us/en/industry/62161/about

[12] Medicare Current Beneficiary Survey (MCBS)

[13] Building Vaccine Confidence Through Community Engagement

[14] Centers for Medicare and Medicaid Services (CMS) Report
https://www.cms.gov/files/document/medicare-covid-19-data-snapshot-services-through-2021-03-20.pdf

[15]
https://www.cms.gov/files/document/medicare-covid-19-data-snapshot-services-through-2021-03-20.pdf