# Identifying members hesitancy to Covid19 vaccination

**Humana-Mays Healthcare Analytics 2021**

# Contents

# 1 Executive Summary

This report focuses on tackling the issue of identifying COVID-19 hesitancy amongst Humana members. The dataset covers the time period from March 2020 to March 2021 and has a binary target variable of 'covid_vaccination'. The target variable reflects if a member has taken the COVID-19 vaccine or not taken.

Firstly, we will discuss the important assumptions we have taken when approaching the data.The first step to approaching any data science problem is to have solid understanding of the data at hand. To build our understanding, we extensively studied existing publications on COVID-19 and vaccine hesitancy within United States and across the globe. We will also be discussing the limitations of the dataset that have been highlighted in the problem statement and how additional information such as contraction of COVID-19 or type of vaccine availability can be crucial at a census level given the nature of the pandemic and how it has evolved since its spread. This assumptions and information will be vital for Humana when formulating an action plan to reach the more vulnerable and under-served population.

Secondly, the report will dive into the analytical section. We will provide an general approach to dealing with data anomalies, the common data cleaning techniques incorporated to ensure minimal data loss and different methods ranging from forward selection algorithm to BorutaPy that were used in feature selection. The cleaned dataset was analysed using Tableau to give a better understanding of the spread and possible biases in the data.

Next, we will highlight the different models that the team ran and tuned different ensemble methods of LightGBM, XGBoost and Gradient-boost. The best scores was produced by the XGBoost model with a Receiver Operating Characteristic(ROC) Area Under Curve(AUC) of 0.6830 and accuracy of 0.8542. We trained the models through a 10 fold cross validation to reduce biases and by maximizing the ROC-AUC curve.

From the model, we extracted the most important features and narrowed them down to based on research and present pandemic situation to formulate actionable insights for Humana. Subsequently, the team concluded that the following factors will be vital to detect Covid_19 hesitancy:

1. Risk adjustment factor(cms_risk_adjustment_factor_a_amt)
2. Reason for entry into Medicare(cms_orig_reas_entitle_cd)
3. age(est_age)
4. trend of cost per month of prescriptions related to VACCINES drugs(rx_gpi2_17_pmpm_cost_t_12-9-6m_b4)
5. late to medicare index(cons_ltmedicr)
6. household food insecurity (%, three-year average), 2013-15(atlas_foodinsec_13_15)

.
Based on this features and the constantly evolving nature of the pandemic, the team decided to formulate a plan that was not solely determinant on the data as that would limit it to just March 2021 when in reality the situation is much more different now with numerous government incentives and possible restrictions on unvaccinated individuals. As the problem is extremely dynamic,the two viable solutions we felt could would work would be:

- Identify members reason to enter medicare and target locations with food insecurity
- Identifying prescription patterns related to common vaccine drug to understand purchase reason and educate on vaccination benefit if needed

The solutions we formulated are more focused at the societal improvement rather than any cost savings for Humana. The inability to identify concrete cost savings is limitation the team identified as existing research in this area is limited and contradicting at some levels.

# 2 Introduction

## 2.1 Understanding the COVID-19 landscape

The COVID-19 pandemic has been a crucial moment in today's digital and highly globalized world. For decades, we were proud of the our advancement as a society, enjoying the fruits of high inter connectivity and endless information at our fingertips. There are always two sides to a coin and COVID-19 made this divide even more apparent.

Information about the pandemic varied vastly across the different continents and in the case of United States within states. Numerous studies have examined the effect of COVID-19 misinformation on public perceptions of the pandemic, the tendency of certain sociopolitical groups to believe misinformation and compliance with public health guidance, including willingness to accept a COVID-19 vaccine. [1].

As of Q2 2020 with the build to the 2020 presidential election, the public's openness to take a vaccine had not been static. It was highly volatile and fluctuated with new information, sentiments swayed around the different COVID-19 vaccines, as well as the general state of the epidemic and perceived risk of contracting the disease. A poll conducted in September 2020 showed significant fall in willingness to accept a COVID-19 vaccine among different genders, all age groups, all ethnicities and all major political groups, very likely due to the heavy politicization of the vaccination in the run up to the presidential election on both sides of the political debate. [2, 3]

Currently there is no solid quantitative assessment of how exposure to misinformation affects intent to receive the vaccine and its implications for obtaining herd immunity if countries adopt this vaccination strategy, a topic greatly debated during the time-span of our dataset .[4]. 2020 really shed a light into how political affiliations, sources of news and social media play an integral role in shaping the cognitive and decision making process of humans during extremely crucial moments. It is essential to understand how misinformation vastly impacts socio-demographic groups and whether groups that are at high risk of developing severe complications from COVID-19 are more vulnerable to this misinformation. [5]

The concept of echo chamber pops up regularly in research study around social media usage and implications. Echo chamber is an environment where a person only encounters information or opinions that reflect and reinforce their own. Echo chambers can create misinformation and distort a person's perspective so they have difficulty considering opposing viewpoints and discussing complicated topics. They are fueled in part by confirmation bias, which is the tendency to favor information that reinforces existing beliefs. Not so surprisingly, consumption of content about vaccines is dominated by the echo chamber effect and the polarization has increased over the years. Majority of users consume information in favour or against the vaccines, not both. [6]

## 2.2 Case Background & Assumptions

Humana is a leading healthcare company that offers a variety of insurance products and health & wellness services. Throughout the COVID-19 pandemic, Humana has worked to overcome challenges in health care delivery in order to provide maximum support for its members' health and well-being. The organization has sent reusable face masks to it Medicare Advantage (MA) and Prescription Drug Plan (PDP) members, delivered over one million preventive care colon cancer screening and diabetic condition management in-home test kits to eligible members, and proactively called and scheduled approximately 65,000 members in 46 states to receive COVID-19 vaccinations, with a focus on individuals with societal barriers to scheduling and receiving the vaccine.

Furthermore, the Centers for Medicare and Medicaid Services (CMS) announced on October 8,2021 that Humana has received a 5 out of 5-star rating for four of its contracts for 2022 and 4.5-star rating for eight Medicare Advantage contracts offered in 33 states. This star rating are a reflection of the company's strong focus on ensuring high quality of care, patient-centered clinical results and reliable member support for its member, especially through this uncharted pandemic times.

The main business objective is to create a model to predict which members are likely to be hesitant so that Humana can design targeted outreaches for these members, prioritized to reach the most vulnerable and under-served populations to receive health solutions.
Given this information, we can assume that Humana has extensively researched and implemented outreach programs' to promote and assist in COVID-19 vaccination amongst it members. The chart below from the Centers for Disease Control and Prevention(CDC) clearly shows how the vaccination uptake has dramatically increased from March 2021, the point where our data ends.

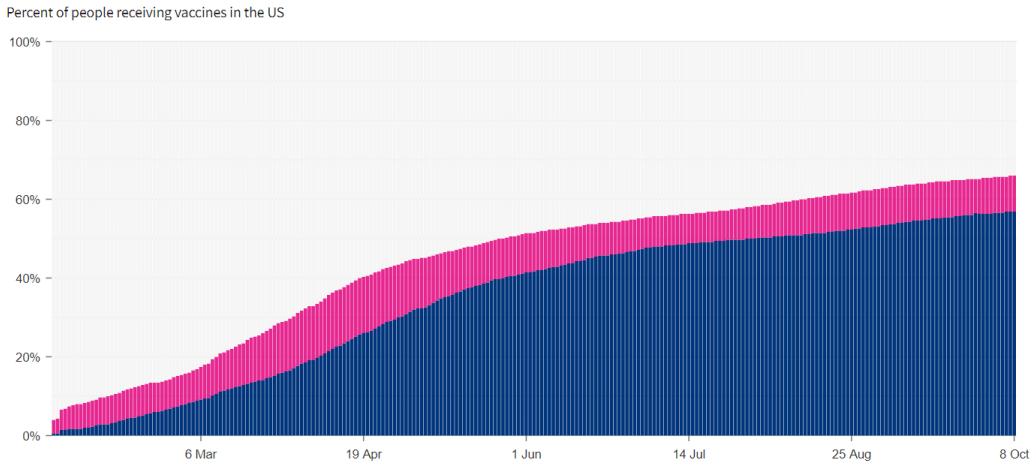


Figure 1: Vaccination Rate in United States

The pink area shows that as of October 8, that 66% of the population have received at least one dose while the blue area indicates that 57% of the population have been fully vaccinated. This is a huge improvement from the 17% one dose and 9% fully vaccinated numbers from March 2021.

The team felt that this development was vital to understand when dealing with historical data from a period where there was speculations about the vaccination,efficacy of the vaccinations and to large extent even the existence of such pandemic.[7]. Furthermore, when we explore the data in the upcoming sections, we will constantly refer to the existing statistics from CDC and literature from the section above to choose a direction in working towards a tangible and feasible solution.

Therefore, with the current information from Humana, CDC and existing literature review, the team made a few key assumptions in relation to COVID-19 hesitancy during the period of March 2020 to March 2021(past) vs March 2021 to October 2021(present). The assumptions are:

- Vaccine hesitancy in the past was primarily due to its novelty
- Geopolitical influences especially during the presidential campaign period affirmed certain subconscious biases
- Rapidly evolving state of the virus and varying WHO information affected confidence
- Different vaccine provider and supply & demand factors played a role
- Stability in vaccinated countries across the world positively affirmed vaccination
- Government incentive to promote vaccination at present played major role in uptake
- With Humana aiding 65,000 members in vaccination matters, the remaining members are the ones with preconceived bias towards vaccines, have already contracted COVID-19 or want to take the vaccine but have an irrational fear due to economic or medical uncertainty.

.
Therefore, the solution proposed later use information from the model to aid in the formulation of media targeting for educational outreach and positive affirmation for the unvaccinated population on how Humana, their insurance provider will be with them throughout their vaccination journey.

# 3 Data Exploration

## 3.1 O in OSEM framework - Obtaining the data

The dataset has provided to us by Humana contained information from multiple databases; ranging from their own internal database of members & claims records to information from USDA at a county/ zip-code level and externally purchased data which provide estimates at individual levels and geographical attributes.

## 3.2 S in OSEM framework - Scrubbing the data

The most important portion of any data science project will be the scrubbing or cleaning of the dataset before performing any sort of analysis on it. We were provided with two datasets, one being the testing dataset and the other was the holdout dataset on which we would test our final model.
The training dataset consisted of 974842rows and 366 columns(excluding index). The dataset was broken into the following sections as detailed in the case presentation deck.

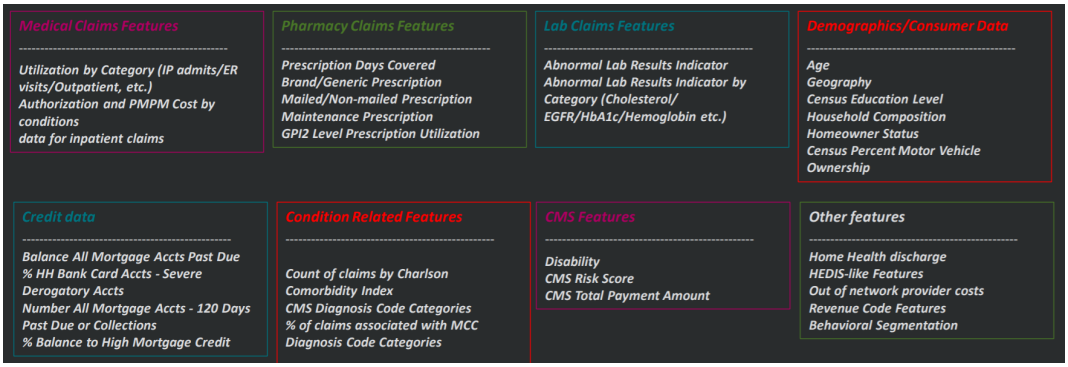


Figure 2: Information from case deck.

With the information from above, we further segregated the data using the prefix from their column names and formed 16 unique groups containing multiple columns and 17 unique columns(including the target variable) that did not belong to any of this groups.

| | |
|---|---|
| lab | lab_albumin_loinc_pmpm_ct |
| lab | lab_dist_loinc_pmpm_ct |
| mcc | mcc_ano_pmpm_ct_t_9-6-3m_b4 |
| mcc | mcc_chf_pmpm_ct_t_9-6-3m_b4 |
| mcc | mcc_end_pct |
| med | med_ambulance_coins_pmpm_cost_t_9-6-3m_b4 |
| med | med_ip_snf_admit_days_pmpm |
| med | med_outpatient_deduct_pmpm_cost_t_9-6-3m_b4 |
| med | med_outpatient_mbr_resp_pmpm_cost_t_9-6-3m_b4 |
| med | med_outpatient_visit_ct_pmpm_t_12-9-6m_b4 |
| med | med_physician_office_allowed_pmpm_cost_t_9-6-3m_b4 |
| med | med_physician_office_ds_clm_6to9m_b4 |
| rej | rej_days_since_last_clm |
| rej | rej_med_er_net_paid_pmpm_cost_t_9-6-3m_b4 |
| rej | rej_med_ip_snf_coins_pmpm_cost_t_9-6-3m_b4 |
| rej | rej_med_outpatient_visit_ct_pmpm_t_6-3-0m_b4 |
| rej | rej_total_physician_office_visit_ct_pmpm_0to3m_b4 |
| rev | rev_cms_ansth_pmpm_ct |
| rev | rev_cms_ct_pmpm_ct |
| rev | rev_pm_obsrm_pmpm_ct |
| rwjf | rwjf_air_pollute_density |
| rwjf | rwjf_inactivity_pct |
| rwjf | rwjf_income_inequ_ratio |
| rwjf | rwjf_men_hlth_prov_ratio |
| rwjf | rwjf_mental_distress_pct |

Figure 3: Grouping based on prefix within the dataset.

Based on this grouping above, we analysed the individual columns to make sense of the information provided within them and we found certain glaring issues that needed to be addressed.

- The data was highly imbalanced with the target variable in the ratio of 83% not having the vaccination and only 17% being vaccinated. This was reflective of the vaccination rate within United States during that period.(Fig 1).
- The race variable was extremely skewed towards the Race_CD:1(White) which comprised almost 83% of the data while the remaining 6 races summed up as the remaining 17%.
- The age spread from 20 to 104. Majority of the data was centralized around 60 to 80 years old as the dataset contained mainly MAP members and also Federal government stipulation around on who could take vaccine during the period the dataset was collected.
- There were 133 columns with 99% of the data in the column only containing a 0. The server lack of variation makes this columns redundant to any machine learning algorithm. Thus this columns were dropped.
- Out of the 233 columns (366-133) 103 columns had missing value ranging from 32.1% to 0.01%. Dropping this columns would result in approximately 44% loss in data. Therefore necessary imputations were carried out such as imputating numerical columns with either mean or median values based on analysis of their histograms and values and including a 'Category_X for columns with missing values.
- Multiple of the 133 columns consisted of '*' variable that either indicated missing data or no information. Dropping this columns resulted only in a 0.4% loss in the data. Given the size of the dataset, this was acceptable.
- 3 out of 233 columns consisted of extremely skewed data; Skewness greater than 50. Transforming this data to fit them into Gaussian distribution for intial models reduced accuracy significantly, therefore they were dropped after in subsequent iterations
- 27 out of 233 columns consisted of relatively skewed data; The columns were transformed using log transformation to fit them into assumed Gaussian distribution

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 972006 entries, 0 to 972005
Columns: 233 entries, atlas_pct_laccess_child15 to race_cd
dtypes: float64(150), int64(14), object(69)
memory usage: 1.7+ GB
```

Figure 4: Details on clean data.

After the numerous cleaning steps, we arrived at a clean training data with 972006 rows and 233 columns which was used for Exploratory data analysis (EDA). The clean dataset contained 150 float columns, 14 integer columns and 69 objects.

## 3.3 E in OSEM framework - Exploring the data

Using the Pandas Profiler feature in Python and after tuning certain hyper-parameters we were quickly able get a detailed breakdown of the different variables.
For the scope of this report in this section, we will focus on few of the factors identified at the start of the report. We will look at how each of them interacts with the target variable.

### 3.3.1 Age

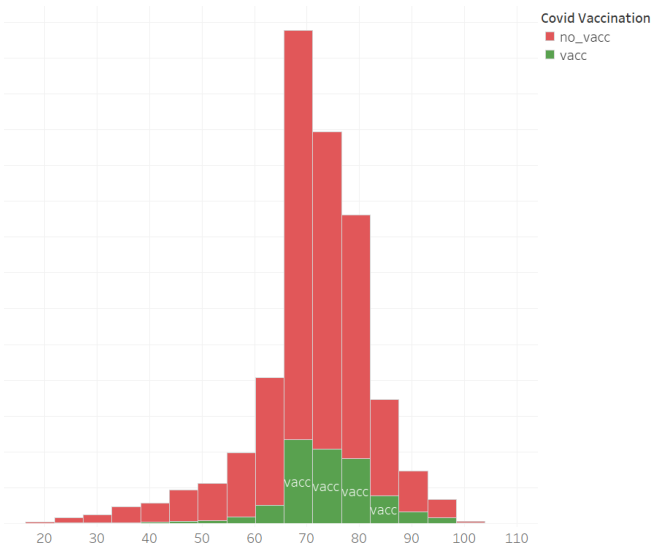


Figure 5: Distribution of vaccinated members by age.

Within the age we observed that the vaccinated individuals were coming from the older age group of around 60-80.This numbers made sense as the first vaccination roll out in United States started on December 14th with priority given to front line workers, high risk individuals and individuals aged 65 above [8].

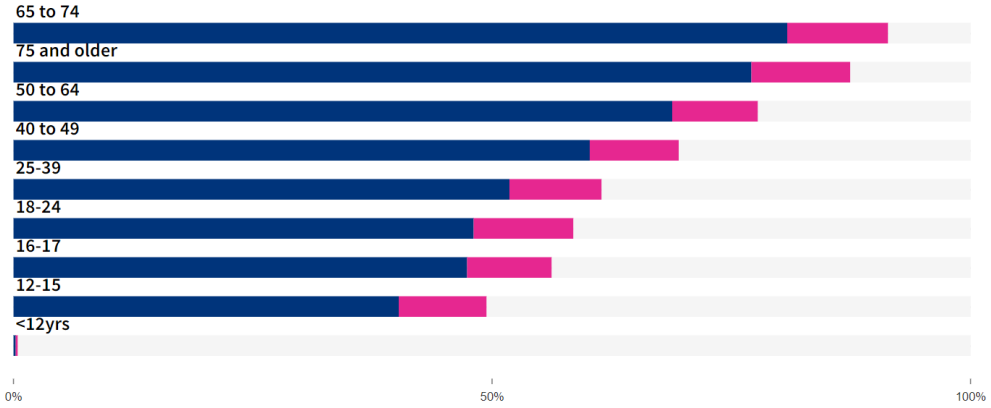


Figure 6: Vaccination Rate by age

This was critical information as it would been wrong to merely concluded that younger age group are not taking the vaccine and also state that the red portion within the higher age group is the ones we should solely target at present. Below is the existing statistic from CDC dated October 8 that states the current vaccination numbers by age group. Additionally, given the spread of the age, we will take into consideration how CDC has classified the age brackets and incorporate the same during our feature engineering.

### 3.3.2 Race

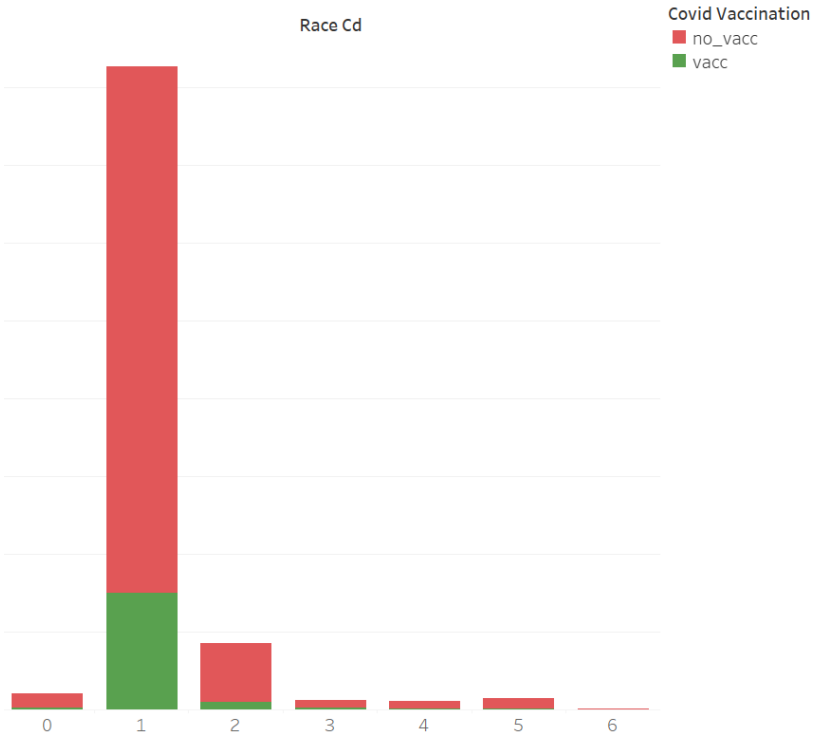


Figure 7: Vaccination Rate by age

Within the race, we can see that the data is highly skewed towards race 1(whites). Numerous studies highlight the Racial and socioeconomic disparity exists in the levels of knowledge, attitude and practices related to COVID-19.[9]. While the aim of the model is to generate actionable solution from the given dataset, we feel that it will be vital in the long run to have instances of other races as it will definitely allow for a more robust and applicable outcome.

## 3.4 M in OSEM framework - Modelling the Data

### 3.4.1 Stratified Test-Train Split

As we observed in the previous sections on how the data is skewed towards a certain race, age and vaccination status.

```
df1["stra_col"] = df1["age_bracket"].astype(str) + "_" + df1["race_cd"].astype(str) + "_" + df1["covid_vaccination"].astype(str)
```

Figure 8: Stratified Column

The best approach to ensure that the test train split did not result in bias towards a certain class was to incorporate a stratified test-train split based on the combinations of variables below into a single columns as such. The column was dropped after the split and there was not data imbalance after the split ensuring the model was fed population representative data.

- Binned Race_1 as 'White' column and the remaining as 'Other' .
- Created age buckets in line with CDC's reporting of 75 and older,65 to 74, 50-64 40-49,39 below(given small number in dataset).
- Used the covid_vaccination status as it is for the new column. .

### 3.4.2 Feature Engineering

Feature engineering is the process of transforming the clean raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. We had already cleaned the data to remove columns with low variation, rows with *(ambiguous) values, scaled the numerical columns with logarithmic transformation and imputed Category X for categorical columns with missing value.

The step for featuring engineering involved instantiating a column transformer object with a StandardScaler object for the numerical columns and OrdinalEncoder and OneHotEncoding for the categorical column. We avoided using OneHotEncoding on all the categorical variables due the curse of dimensionality. We investigated all the categorical columns using a user defined function to extract the unique values and mapped it back to the column descriptions. Categorical columns that displayed an ordered were categorically encoded and the remaining that did not have a distinct order in their description were OneHotEncoded. This column transformer object was then used to transform the testing dataset to prevent data leakage.

### 3.4.3 Feature Selection

Once the data had been scaled and transformed appropriately, tested for correlation amongst the variables with a threshold of 70% and removed several variables that were highly correlated.

Lastly, we used forward, backward and BorutaPy feature selection methodologies for feature selection. Analysing the outputs of each feature selection method, the BorutaPy wrapper algorithm fitted the XGBoost resulted in a total of 56 features that logically made sense based on existing research about vaccine hesitancy.

| | | | |
|---|---|---|---|
| 'credit_hh_nonmtgcredit_60dpd', | 'rev_pm_obsrm_pmpm_ct', | 'rx_generic_mbr_resp_pmpm_cost', | 'rx_generic_pmpm_cost_6to9m_b4', |
| 'rx_bh_pmpm_ct_0to3m_b4', | 'atlas_pct_sfsp15', | 'lab_dist_loinc_pmpm_ct', | 'rx_nonbh_pmpm_ct_0to3m_b4', |
| 'rx_overall_gpi_pmpm_ct_0to3m_b4', | 'cms_tot_partd_payment_amt', | 'atlas_pct_nslp15', | 'rx_tier_1_pmpm_ct_0to3m_b4', |
| 'zip_cd', | 'cons_cgqs', | 'rx_gpi2_56_dist_gpi6_pmpm_ct_3to6m_b4', | 'atlas_hipov_1115', |
| 'credit_bal_consumerfinance', | 'rx_nonbh_mbr_resp_pmpm_cost', | 'rx_generic_pmpm_ct_0to3m_b4', | 'cons_estinv30_rc', |
| 'atlas_vlfoodsec_13_15', | 'rx_days_since_last_script', | 'rx_overall_mbr_resp_pmpm_cost_0to3m_b4', | 'atlas_pct_sbp15', |
| 'credit_hh_bankcard_severederog', | 'rwjf_uninsured_child_pct', | 'rx_tier_2_pmpm_ct_3to6m_b4', | 'atlas_pct_cacfp15', |
| 'cnt_cp_webstatement_pmpm_ct', | 'atlas_net_international_migration_rate', | 'rx_maint_pmpm_ct_9to12m_b4', | 'pdc_lip', |
| 'rwjf_uninsured_adults_pct', | 'atlas_foodinsec_child_03_11', | 'cms_risk_adjustment_factor_a_amt', | 'rx_tier_2_pmpm_ct', |
| 'lab_albumin_loinc_pmpm_ct', | 'cons_nwperadult', | 'rx_generic_pmpm_cost', | 'atlas_foodinsec_13_15', |
| 'race_cd', | 'est_age' | 'cms_orig_reas_entitle_cd', | 'atlas_type_2015_mining_no', |
| 'cons_rxadhm', | 'sex_cd', | 'hum_region', | 'cons_stlnindx', |
| 'rx_phar_cat_cvs_pmpm_ct_t_9-6-3m_b4', | 'cons_rxadhs', | 'rx_mail_net_paid_pmpm_cost_t_6-3-0m_b4', | 'mcc_chf_pmpm_ct_t_9-6-3m_b4', |
| 'cons_hxmioc', | 'rx_gpi2_17_pmpm_cost_t_12-9-6m_b4', | 'cons_ltmedicr', | 'cons_lwcm07', |

Figure 9: BorutaPy Columns

Boruta does not compare the importance of a feature with other features. Instead, each feature competes with the 'shadows' of other features which are randomized version of them. Moreover, an important advantage of Boruta over other algorithms is that it provides a statistically robust threshold for selecting important features rather than picking a value subjectively

### 3.4.4 Testing Different Models

We tested three models given the :

- LightGBM - A gradient boosting framework that uses tree based learning algorithms and is designed to be fast and efficient with lower memory usage, better accuracy, and capable of handling large-scale data through parallelization.
- Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. Errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error and it is used to minimize bias error of the model.
- XGBoost - An optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework.

For each of the model we incorperated the following sequential procedure:

1. Hyperparameters tuning - Parameters which define the model architecture are referred to as hyperparameters and thus this process of searching for the ideal model architecture is referred to as hyperparameter tuning. We incorporated a grid search approach which is one of the simple yet most effective hyperparameter tuning techniques. The downside to this brute-force grid search approach with a inclusion of 10 fold cross validation step is that given the size of the dataset computational power on the local system becomes a bottleneck it takes an extremely long time . We explored the option to use the Apache Spark framework on Databricks through Google Cloud Platform (GCP) to utilize the parallel processing to speed up the model evaluation time. However, the team did not want to incur additional cost due to the number of parameters settings defined and thus we decided to reduce the number of cross validation folds to 3 and run the models on our local machines.
2. Fitting the model - For each model once the best hyperparameters were identified, it was used to fit the model on the training dataset.

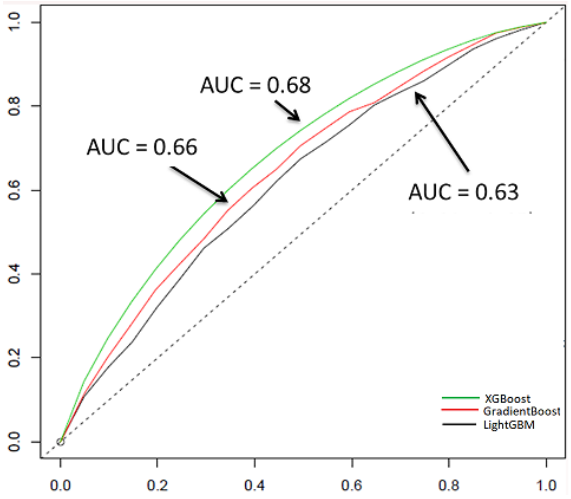


Figure 10: ROC-AUC Curve

3. Mode Evaluation.- Models that generate ROC curves that are closer to the top-left corner, give larger AUC than the baseline, and therefore, they are performing better. Because the baseline model is operating based on the Bernoulli distribution with p = 1/2, there is a 50% chance that the true class of each member is identified correctly. Therefore, the recall score for each class is 0.5. We compare our model performance with this baseline. We can see that the XGBoost gives the best ROC-AUC score amongst the three models tested.

.

### 3.5 Understanding the Model - Shapely Additive Explanations

There are many various strategies to improve your model understanding, and one of them is to emphasize the importance of features. Using feature importance, you may calculate how much each feature of your data contributes to the model's prediction. You can determine which attributes have the greatest impact on your model's decision-making by conducting feature importance tests. You can act by deleting elements that have a minor impact on the model's predictions and concentrating on improving the more important features. This has the potential to greatly increase model performance.

There are numerous methods for determining the relevance of a characteristic. Stat models and scikit-learn are used in some of the basic procedures. Shapely Additive Explanations (ShAP) is a new method employed. Because many of these methods can be inconsistent, the most significant features may not always be given the highest feature priority score, this method is deemed slightly better than typical scikit-learn methods.

The Shapley value explanation is depicted as an additive feature attribution approach, a linear model, which is one of SHAP's innovations. LIME and Shapley Values are linked in this way. According to SHAP, the explanation is as follows:

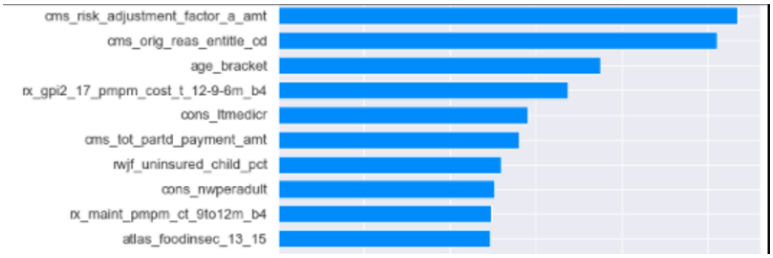


Figure 11: Top 10 Shap features

For a given member, m, the summation is over all the subsets S, of the team, T=1,2,3,. . . ,p,that one can construct after excluding m. In the above formula, k(S) is the size of S, v(S) is the value achieved by sub team S, and v(Sm) is the realized value after m joins S.
We define a fair allocation to be an allocation that satisfies:

- Efficiency: The total of individual contributions is equal to the team's realized value:

$$\sum_{m=1}^{m=p} \varphi_m(v) = v(T)$$

Figure 12: Equation

- Symmetry: Two team members with the same added value have the same share:

if $v(S \cup \{m\}) = v(S \cup \{n\})$, then $\varphi_m(v) = \varphi_n(v)$

Figure 13: Equation

- Linearity: If the team participates in several projects (say two projects), each yielding v(T), u(T), then adding the share of each member in the different projects is the same as finding his/her share using the total gain v(T)+u(T). In other words, the shares are additive(Figure 14)

  To assess the relevance of feature j, think of the process as drawing feature values in random order for all features except feature j for each iteration, then calculating the difference between prediction with and without feature j. The average difference from all combinations is used to

$$\varphi_m (v + u) = \varphi_m (v) + \varphi_m (u) \text{ And, } \varphi_m (av) = a\, \varphi_m (v).$$

Figure 14: Equation

calculate the Shapley value. The Shapley value is the average marginal contribution of a characteristic when all feasible combinations are considered.

The key benefit of the Shapley value is that it contributes a significant number of features with mathematically established theory and properties concerning its mechanism.

By virtue of this method, the top 40 features that has high impact on the dependent variable is procured. To mention some, features like "Member age", "Risk Adjustment Factor A Amount", "Code indicating the original reason for entry into Medicare" have high impact, while features like "count per month of prescriptions related to Tier 2 drugs in the past third to sixth month prior to the score date", "claims per month for a revenue code related to specialty services in the past one year", "member responsible cost per month of prescriptions related to generic drugs in the past one year" are the extremities having lesser importance but might have indirect impact on the model.

### 3.6 Tentative Solution

Socio-economic vulnerability refers to an individual's, a group's, or a community's capacity to cope with and adapt to external stressors affecting their livelihoods and well-being. Vulnerability has arisen as a helpful and contentious notion for comprehending, quantifying, and evaluating people's vulnerability to any catastrophe, here Covid-19 vaccine hesitancy. Both resilience and vulnerability are ambiguous concepts, particularly because robust systems are frequently assumed to be less vulnerable than non-resilient systems.

The idea of community vulnerability is viewed analytically in this work as encompassing demographical features, socio-economic status, healthcare availability, insurance availability and epidemiological factors. These six categories are created by grouping the top 22 features which showed high importance in shapely feature selection. provides information on the vulnerability index category used and description of indicators
.
The ordinal vulnerability rankings of low (1), moderate (2), and high (3) were classified based on the Social Vulnerability Index (SVI). Each feature was deemed to have a direct or indirect effect on community vulnerability, and the following impact factors were assigned to each indication: Direct impact is equal to 1, whereas indirect impact is equal to 0.5. The following equation was used to determine the level of vulnerability for each category
.

$$\text{SVI} = \sum_{k=0} ka$$

Figure 15: Equation

Where a is the impact factor (direct impact = 1; indirect impact = 0.5) and k = 2 ((1+2+3)/3)), where k indicated to the three vulnerability categories.

| Vulnerability Category | Indicator |
|---|---|
| Demographical Features | Member age |
| | Prescriptions related to hyperlipidemia in the past one year |
| | Ethnicity |
| | Member gender |
| | Member zip code |
| | Household food insecurity |
| | Child & Adult Care |
| | School Breakfast Program participants |
| Socio-Economic Status | Census Geo-unit Quality Score |
| | Estimated Household Investable Assets Recoded |
| | Net Worth Per Adult |
| | Balance Non-Mortgage Loan Accts 60+ Days Past Due |
| | Household very low food security |
| Healthcare Availability | Code indicating the original reason for entry into Medicare |
| | Risk Adjustment Factor A Amount |
| | Late to Medicare Index |
| | Total Part D Payment Amount |
| Insurance Availability | Clinical Care - % adults under age 65 without health insurance |
| | Clinical Care - % children under age 19 without health insurance |
| Epidemiological Factors | cost/month of prescriptions related to generic drugs in the past one year |
| | trend of cost per month of prescriptions related to VACCINES drugs in the past sixth to ninth month versus ninth to twelfth month prior to the score date |
| | count per month of prescriptions related to maintenance drugs in the past ninth to twelfth month prior to the score date |

Figure 16: Key vulnerability indicators

### 3.6.1 Social Vulnerability Index result

For each of the six groups, a vulnerability score has assigned based on the vulnerability indices employed. The vulnerability score of important vulnerability indicators is shown in Table below, along with their impact on the community.

.

For each of the six groups, a vulnerability score has assigned based on the vulnerability indices employed. The vulnerability score of important vulnerability indicators is shown in Table below, along with their impact on the community. The result shows the social vulnerability for the six categories. Socio-economical status shows the highest vulnerability score of 26. Demographical features, health care availability and epidemiological factors showed similar vulnerability towards the vaccine hesitancy.

.

Insurance availability has the least SVI. Socio-economic status has more indicators since there were more data availability towards this category. Though being the least vulnerable category, all indicators of the Insurance availability exhibit high impact factor. Socio-economical status contains attributes that have at most significance towards the Covid-19 vaccination hesitancy, which was also evident from the shapley feature importance. To mention some features such as food insecurity, child and adult care, geo-unit quality score and net worth have exerted adverse effect on the Covid-19 vaccination hesitancy.

| Vulnerability Category | Indicator | Impact Factor | K Value | SVI |
|---|---|---|---|---|
| Demographical Features | Member age | 1 | 4 | 12 |
| | Prescriptions related to hyperlipidemia in the past one year | 0.5 | 2 | |
| | Ethnicity | 0.5 | 2 | |
| | Member gender | 0.5 | 2 | |
| | Member zip code | 0.5 | 2 | |
| Socio-Economic Status | Household food insecurity | 1 | 4 | 26 |
| | Child & Adult Care | 1 | 4 | |
| | School Breakfast Program participants | 0.5 | 2 | |
| | Census Geo-unit Quality Score | 1 | 4 | |
| | Estimated Household Investable Assets Recoded | 0.5 | 2 | |
| | Net Worth Per Adult | 1 | 4 | |
| | Balance Non-Mortgage Loan Accts 60+ Days Past Due | 0.5 | 2 | |
| | Household very low food security | 1 | 4 | |
| Healthcare Availability | Code indicating the original reason for entry into Medicare | 1 | 4 | 12 |
| | Risk Adjustment Factor A Amount | 0.5 | 2 | |
| | Late to Medicare Index | 1 | 4 | |
| | Total Part D Payment Amount | 0.5 | 2 | |
| Insurance Availability | Clinical Care - % adults under age 65 without health insurance | 1 | 4 | 8 |
| | Clinical Care - % children under age 19 without health insurance | 1 | 4 | |
| Epidemiological Factors | cost/month of prescriptions related to generic drugs in the past one year | 1 | 4 | 12 |
| | Trend of cost per month of prescriptions related to VACCINES drugs in the past sixth to ninth month versus ninth to twelfth month prior to the score date | 1 | 4 | |
| | Count per month of prescriptions related to maintenance drugs in the past ninth to twelfth month prior to the score date | 1 | 4 | |

Figure 17: Equation

This section conducted an in-depth analysis of community-level susceptibility to Covid-19 vaccination hesitancy. By creating variables for each of the six vulnerability index categories. This study established a practical and analytical framework for assessing the vulnerability of indicators. Using the framework of each categories, the socio-economical factors bring out more hesitancy towards the Covid-19 vaccine.

# References


[1] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199, 2020.

[2] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[3] Patrick Peretti-Watel, Valérie Seror, Sébastien Cortaredona, Odile Launay, Jocelyn Raude, Pierrea Verger, Lisa Fressard, François Beck, Stéphane Legleye, Olivier l'Haridon, et al. A future vaccination campaign against covid-19 at risk of vaccine hesitancy and politicisation. *The Lancet Infectious Diseases*, 20(7):769–770, 2020.

[4] Daniel Romer and Kathleen Hall Jamieson. Conspiracy theories as barriers to controlling the spread of covid-19 in the us. *Social science & medicine*, 263:113356, 2020.

[5] Pascal Geldsetzer. Knowledge and perceptions of covid-19 among the general public in the united states and the united kingdom: a cross-sectional online survey. *Annals of internal medicine*, 173(2):157–160, 2020.

[6] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. Polarization of the vaccination debate on facebook. *Vaccine*, 36(25):3606–3612, 2018.

[7] Roland Imhoff and Pia Lamberty. A bioweapon or a hoax? the link between distinct conspiracy beliefs about the coronavirus disease (covid-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, 11(8):1110–1118, 2020.

[8] Peter Loftus and Melanie Grayce West. First covid-19 vaccine given to u.s. public, Dec 2020.

[9] Wilson M Alobuia, Nathan P Dalva-Baird, Joseph D Forrester, Eran Bendavid, Jay Bhattacharya, and Electron Kebebew. Racial disparities in knowledge, attitudes and practices related to covid-19 in the usa. *Journal of Public Health*, 42(3):470–478, 2020.