Predicting the Likelihood of Hesitancy to COVID-19 Vaccine for Medicare Members Using a Gradient Boosting Tree Approach

Humana-Mays Healthcare Analytics Case Competition
October 2021

Table of Contents

1 Introduction

1.1 Background

In the past year, COVID-19 had caused a significant impact in our lives. As of September 15th, 2020, there had been more than 6.5 million confirmed COVID-19 cases and more than 195,000 deaths in the United States [1]. The economic crisis is unprecedented in its scale: the pandemic has created a demand shock, a supply shock, and a financial shock all at once. More seriously, it is only the data from 2020 and covid-19 still kept impacting our lives negatively. Fortunately, the COVID-19 vaccine has been developed and put into use since the spring of 2021. According to Yale Medicine, COVID-19 vaccines are really effective in "helping us protect against severe disease and death from variants of the virus that causes COVID-19 currently circulating, including the Delta Variants", and also it suggests that the side effects after vaccination are "normal" and should be gone in a few days [2]. Also, this gave us the hope to defeat COVID-19 completely, by achieving the herd immunity where a large portion of a community (the herd) becomes immune to a disease, making the spread of disease from person to person unlikely [3]. Vaccination is the safest method for disease protection and reaching herd immunity. Vaccines expose your body to a weakened or killed form of the pathogen, triggering your immune system to identify and protect you from future threats [3]. The goal is to slow the spread of disease, build lasting immunity for everyone and protect vulnerable populations without risking lives.

However, due to different reasons, people fail to get injected. As of October 6th, 2021, around 56.5% population in the United States got fully vaccined, which is still much lower than the percentage required for herd immunity [4]. On the other hand, herd immunity doesn't protect everyone who doesn't get vaccinated: deciding to refuse vaccination can leave the person and his/her community vulnerable to getting and spreading a disease, even if the person have antibodies from a previous infection [3].

According to AMA, there is an interesting idea that "the biggest impediment to getting more people fully vaccinated for COVID-19 is access, not vaccine hesitancy [5]. However, today the COVID-19 vaccine is more available so this concern is no longer plausible. So, what we want to do is to estimate those people who are hesitated about the vaccine. According to John Hopkins Medicine, many people are hesitated about the Covid-19 vaccine for the following reasons. For example, they care about the safety of vaccine since it was created very soon. Also, the cared about the potential side effects that were not yet discovered. Moreover, some people cared about the allergy's effects [6]. So, we want to build a model to filter those who hesitate about the vaccine.

1.2 The Humana Competition

According to Humana, Humana had focused on the "most vulnerable and underserved populations". However, there are still some members who are still hesitated about the vaccine for some reasons. So finding out those people and give them personalized conversation is important. We can see similar report from NYTimes that "In dozens of interviews on Thursday in eight states, at vaccination clinics, drugstores and pop-up mobile sites, Americans who had finally arrived for their shots offered a snapshot of a nation at a crossroads — confronting a new surge of the virus but only slowly embracing the vaccines that could stop it". Until someone they trust convinced them.

Based on the demanding global impact of Covid-19 and the urgent need to find out who are hesitated about the vaccine, It is import to use data and potential public data to create a model in order to target them. Therefore, we can take further steps as soon as possible to help them build trust in vaccine and provide possible solutions. In this way, Humana could know the potential customers who would get sick due to the COVID-19. By filtering them and offer them help, Humana could reduce potential medical cost for not paying the insurance. Therefore, Humana could make more money so this model is crucial.

2 Data Preparation

2.1 Analytical tools

All of our analysis was coded in R and we collaborated in Google Colab notebook. For data exploration and data cleaning process we used packages like tidyverse, Hmisc and SmartEDA, and for modeling part we used caret, xgboost, glmnet, pROC and so on.

2.2 Exploratory Data Analysis

2.2.1 Understanding the dataset

In the training dataset that we have, there're 974,842 records(members) with 367 variables. The target variable is *covid_vaccination*: =vaccination if the member was vaccined, =no_vaccination if the member was not vaccined. For analyzing purposes, based on our understanding of the problem, we redefine this column as a binary indicator: =1 if the member was not vaccined, =0 if the member was vaccined. Other features mainly fall into 7 groups:

- Medical claims features
- Pharmacy claims features
- Lab claims features
- Demographics/Customer data
- Credit data

- Condition Related features
- CMS features

And the features are categorized into three types: integer, character and numeric.

2.2.2 Data Cleaning

In this process, we looked into the distribution of each feature carefully and tried to handle each case in order to get more "cleaned" data.

We noticed that there're a lot of missing values in many variables. For different variables we chose to use different imputation methods to fill in those NAs to complete the dataset. For binary and categorial variables we impute with the mode, while for numeric variables we impute with the median of the respective column.

In particular, we found several groups of variables that are partly missing for quite large proportion of members (20% or so). That could be caused by different data sources that Humana might be using. We first filled those missing values with the median, and then added a binary column indicating whether the member belongs to the group that have these variables unavailable.

After that we convert each non-numeric value to numbers so that we can handle them easily. For example, there're some * in the records, we interpret that as no records and then convert to 0.

Moreover, we eliminated some variables by looking at the distribution of the features one by one and find those that show an unstable or useless pattern which could lead to bias to the model.

We consider above as a very important step for data preparation, since the quality of data would significantly influence the performance of the models.

3 Modeling

3.1 Model Selection

After evaluating different models selection, we decided to use XGBoost, a gradient boosting algorithm, for building our predictive model. XGBoost implements a gradient boosting decision tree algorithm designed to be incredibly efficient and accurate. Boosting is an ensemble approach where new models are created based on weak predictors, with their residuals or errors being combined together to make a final prediction. In order to minimize loss, it uses a gradient descent algorithm when adding to the new models.
To maximize performance of our XGBoost model, we utilized a 10-fold Grid Search to tune the models hyperparameters. Our final models hyper parameters were as follows:

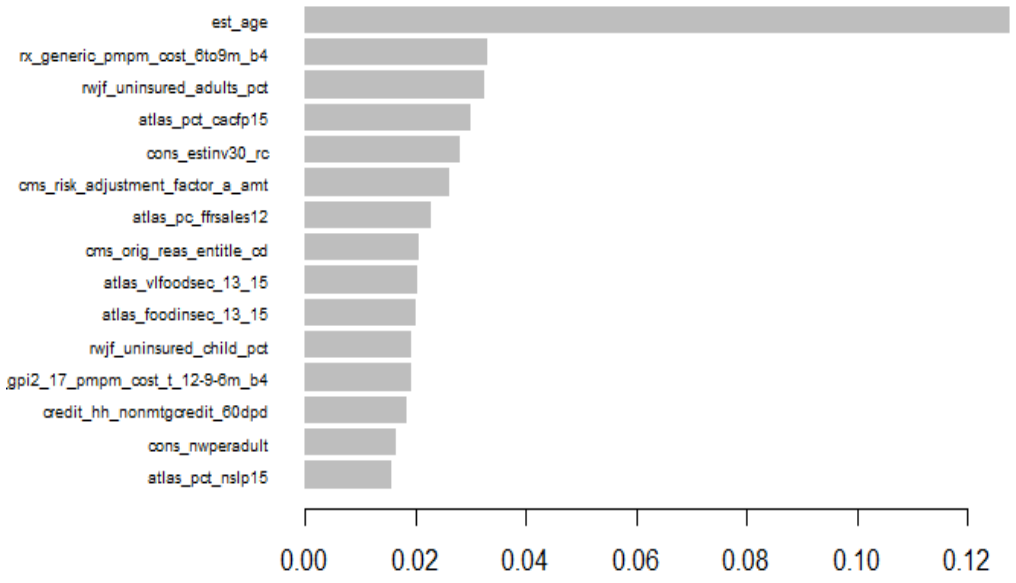
#nthread = 4,

#eta = 0.05,
#max_depth = 7,
#min_child_weight = 5,
gamma = 0,
#subsample = 0.8,
#colsample_bytree = 0.7,
#colsample_bylevel = 0.6,
nrounds = 100

The overall in sample AUC is around 0.67.

3.2 Feature Importance

To better understand which features could be the key indicator in our model, the importance of top 15 features and their descriptions are exported below.



est_age	Member age {calculated using est_bday, relative to score/index date}
rx_generic_pmpm_cost_6to9m_b4	cost per month of prescriptions related to generic drugs in the past sixth to ninth month prior to the score date
rwjf_uninsured_adults_pct	Clinical Care - Percentage of adults under age 65 without health insurance
atlas_pct_cacfp15	Child & Adult Care (% pop)
cons_estinv30_rc	Estimated Household Investable Assets Recoded
cms_risk_adjustment_factor_a_amt	Risk Adjustment Factor A Amount
atlas_pc_ffrsales12	Expenditures per capita, fast food
cms_orig_reas_entitle_cd	Code indicating the original reason for entry into Medicare
atlas_vlfoodsec_13_15	Household very low food security (%, three-year average), 2013-15
atlas_foodinsec_13_15	Household food insecurity (%, three-year average), 2013-15
rx_gpi2_17_pmpm_cost_t_12-9-6m_b4	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 {Based on GPI2 grouping}
rwjf_uninsured_child_pct	Clinical Care - Percentage of children under age 19 without health insurance
credit_hh_nonmtgcredit_60dpd	% HH Non-Mortgage Loan Accts 60+ Days Past Due
cons_nwperadult	Net Worth Per Adult
atlas_pct_nslp15	National School Lunch Program participants (% pop)

4 Recommendations

4.1 Timed and targeted publicity

The feature importance shows that the region of the customers would determine whether they would accept or reject the vaccine in a great extent. Especially for those living in counties with low net income, it is more likely for them to show a hesitancy or even resistance towards COVID-19 vaccines. Therefore, we suggest that Humana could spend some budgets on those counties to convince people to accept vaccines through organized events.

4.2 Differentiated approach based on age

Member age is the most importance key indicator in our model, and it showed that compared with aged, young people could be more likely to reject the vaccine. With extensive research and analysis, we believe that peers could have a better effect in convincing young people in most situations. Therefore, we suggest that Humana could hire contracted young representatives specifically to provide customized approach towards young people. On the other hand, the current employees could focus more than convincing the aged people, which consisted more than 60% of our training dataset.

5 Conclusion

Taking COVID-19 vaccines is importance to win this war between human-beings and virus. In this project, a tree model based on gradient boosting has been built to predict the likelihood of hesitancy to COVID-19 Vaccine for Medicare Members. It could be used for Humana to better approach members and offer help to those really need. For public, this would be beneficial to increase the fully vaccination rate and hence protect more people and their family from damages in health and wealth. It could also help Humana to gain more return of investment by decreasing the payment in claims related to COVID-19. Two recommendations were proposed based on the data analysis: firstly timed and targeted publicity could be conducted, especially in underprivileged counties; secondly differentiated approach according to the age gap could receive better feedbacks.

6 References

[1] L. Bauer, K. Broady, W. Edelberg and J. O'Donnell, "Ten Facts about COVID-19 and the U.S. Economy," *Economic Facts,* p. 2, 2020.

[2] K. KATELLA, "Comparing the COVID-19 Vaccines: How Are They Different?," YaleMedicine, 09 2021. [Online]. Available: https://www.yalemedicine.org/news/covid-19-vaccine-comparison.

[3] V. Iyer, "Herd immunity: What it is and why it's important," AllinaHealth, May 2021. [Online]. Available: https://www.allinahealth.org/healthysetgo/prevent/herd-immunity-what-it-is-and-why-its-important.

[4] Mayoclinic, "U.S. COVID-19 vaccine tracker: See your state's progress," 2021.

[5] B. Murphy, "Access, not hesitancy, now biggest barrier to COVID-19 vaccination," AMA, May 2021. [Online]. Available: https://www.ama-assn.org/delivering-care/public-health/access-not-hesitancy-now-biggest-barrier-covid-19-vaccination.

[6] M. M. Sherita Hill Golden, "COVID-19 Vaccine Hesitancy: 12 Things You Need to Know," 2021. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/covid19-vaccine-hesitancy-12-things-you-need-to-know.