Humana/Mays

2021 Healthcare Analytics

Case Competition

Who is Most Likely to be Resistant to the COVID Vaccination

1. Business Understanding

According to CDC research results, COVID 19-vaccines are safe and effective. They can keep people from getting and spreading the virus that causes COVID-19, and also help keep people from getting seriously ill even if you do get COVID-19, which can protect people's health, reduce possible claims and therefore, reduce the compensation cost of Humana. So it is important to increase the COVID-19 vaccination rates among Humana's members for member and the larger population health and safety and profitability of Humana.

Now, vaccination opportunities are provided to more and more people with lower barriers to be accessed. A study conducted by researchers from Carnegie Mellon University (CMU) and the University of Pittsburgh found that COVID-19 vaccine hesitancy among American adults fell by one-third in the first five months of 2021. However, distrust of vaccines and the government are still keeping many people from getting vaccinated. We believe that these members hesitant or resistant to the COVID vaccination require personalized outreaches involving clinical conversations to build trust in the vaccine. Therefore, our research is conducted to help Humana identify the members who are most likely resistant to the COVID vaccination, and then provide different recommendations and potential solutions to different sub-segments of hesitant members in order to drive vaccination among them.

2. Data Understanding

To achieve our objectives, we are going to use the train data to create a predictive model, then apply it to the handout dataset to predict whether the members in the holdout dataset are willing to vaccinate or not. To understand whether a member accepts COVID vaccination or not, we use the binomial variable “covid_vaccination” from the data dictionary as the dependent variable. If covid_vaccination = 1, it means this member accepts COVID vaccination, while covid_vaccination = 0 means this member is hesitant or resistant to COVID vaccination. After cleaning and selection, other remaining variables in the data dictionary are treated as independent variables.

3. Data Preparation

a. Delete columns

As for columns that only include NA or blank string or 0, we delete these columns as they are not informative. Also, for the character variables that have a large amount of missing values, it's difficult to assign reasonable values for the missing ones and thus we delete the corresponding columns.

b. Delete rows

For character variables with a small number of NAs, we simply drop the rows that contain the NAs.

c. Fill missing values with mean

For numeric variables with missing values, we fill the missing values with the mean of the remaining ones as alternatives.

d. Dummy Variables

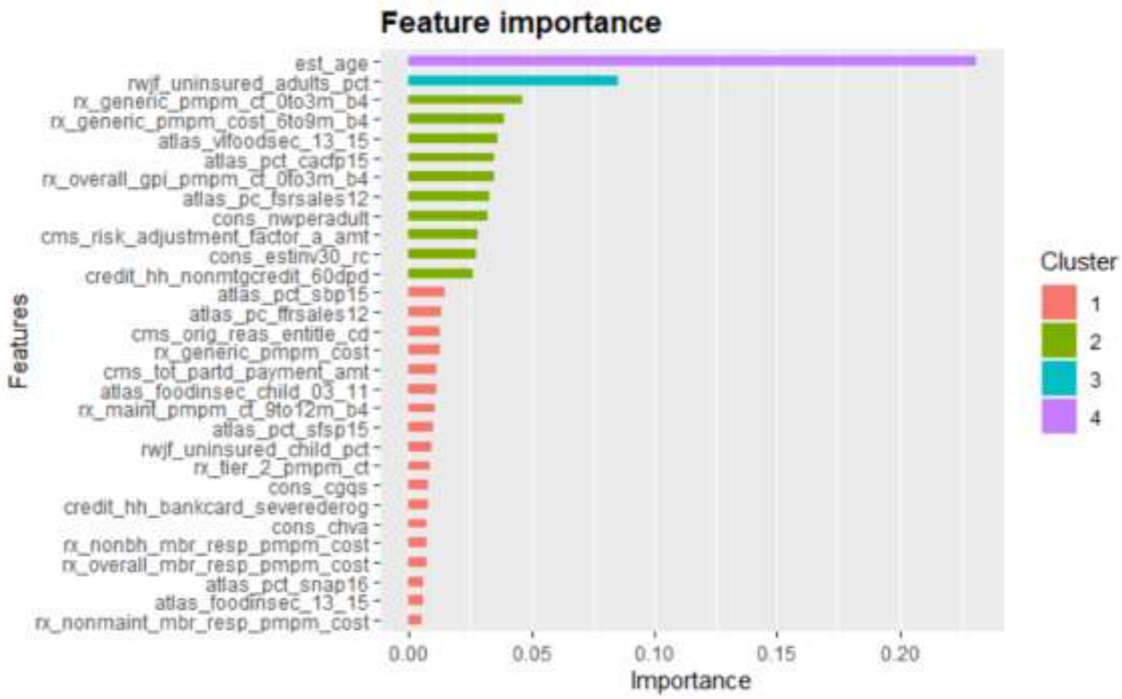
We create several dummy variables to better deal with some character variables, such as sex_cd, covid_vaccination.

4. Modeling

After cleaning the data, we used the cleaned train dataset to create the model. We chose XGBoost as our model. XGBoost uses regularization to reduce variance, and thus avoid overfitting ,which is essential for such a huge dataset we are dealing with.

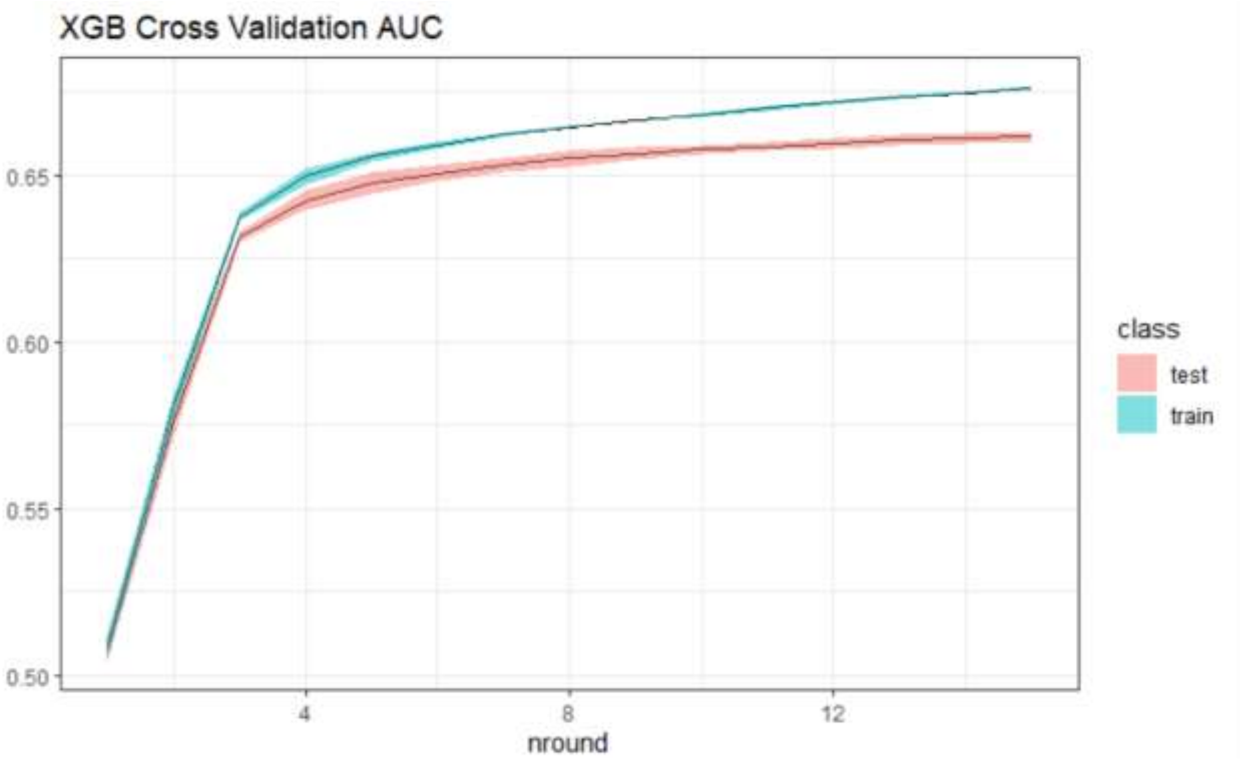
After using the XGBoost model, we identified some important variables as key drivers of our outcome. We can know from the below graph that “est_age”, which means member’s age, is the most important feature influencing people’s acceptance of COVID vaccination. That’s not so surprising because people may think that elderly people and children are more vulnerable so it is more dangerous for them to vaccinate. Also, elderly tend to have more conservative thoughts and may be hesitant or resistant to the COVID vaccination. Another possible reason is that elderly and children have less access to news and information about COVID vaccination.

The second important feature is “rwjf_uninsured_adults_pct”, which means the percentage of adults under age 65 without health insurance. It’s also reasonable for this feature to affect the acceptance of COVID vaccination because health insurance may affect people’s attitude to health and cost of going to the hospital and buying medicine. The next 2 important variables are count per month of prescriptions related to generic drugs in the past three months prior to score date and cost per month of prescriptions related to generic drugs in the past sixth to ninth month prior to the score date, both related to prescriptions about generic drugs. This may be because that people have more convenient ways to have access to the generic drugs.

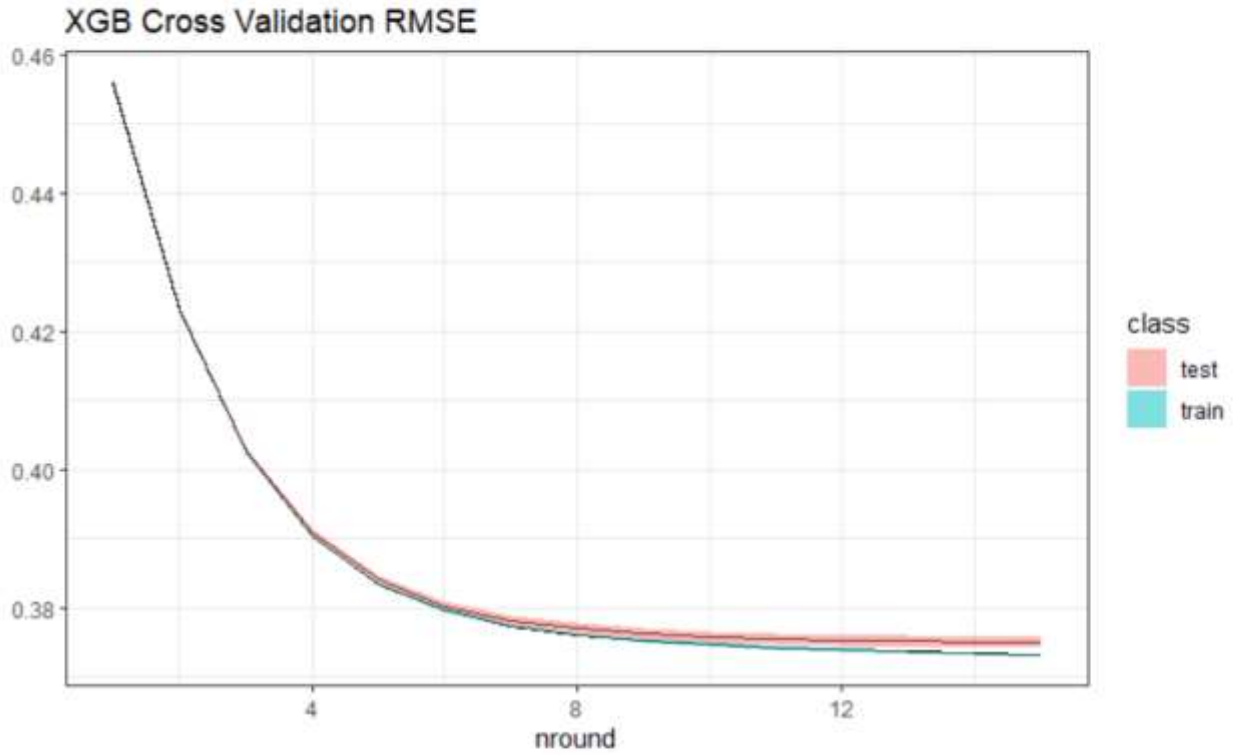

Feature importance
Features
est_age
rwjf_uninsured_adults_pct
rx_generic_pmpm_ct_0to3m_b4
rx_generic_pmpm_cost_6to9m_b4
atlas_vlfoodsec_13_15
atlas_pct_cacfp15
rx_overall_gpi_pmpm_ct_0to3m_b4
atlas_pc_fsrsales12
cons_nwperadult
cms_risk_adjustment_factor_a_amt
cons_estinv30_rc
credit_hh_nonmtgcredit_60dpd
atlas_pct_sbp15
atlas_pc_ffrsales12
cms_orig_reas_entitle_cd
rx_generic_pmpm_cost
cms_tot_partd_payment_amt
atlas_foodinsec_child_03_11
rx_maint_pmpm_ct_9to12m_b4
atlas_pct_sfsp15
rwjf_uninsured_child_pct
rx_tier_2_pmpm_ct
cons_cgqs
credit_hh_bankcard_severederog
cons_chva
rx_nonbh_mbr_resp_pmpm_cost
rx_overall_mbr_resp_pmpm_cost
atlas_pct_snap16
atlas_foodinsec_13_15
rx_nonmaint_mbr_resp_pmpm_cost
0.00
0.05
0.10
0.15
0.20
Importance
Cluster
1
2
3
4

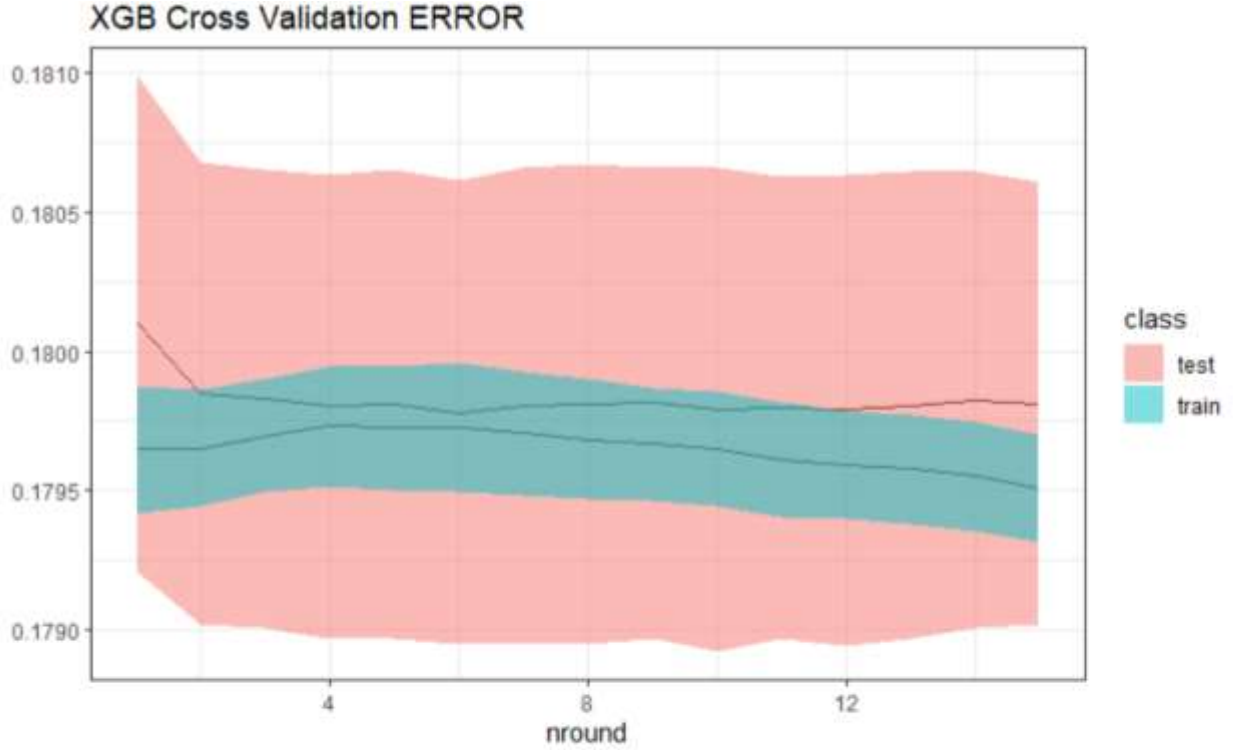
5. Evaluation

From the results of the XGBoost model, we can know that the AUC of the train data is about 0.675, and the AUC of the test data is about 0.66, which are both higher than 0.5. Therefore, we can conclude that our model's accuracy is much better than the baseline model.



Also, our RMSE is about 0.37, and error is less than 0.18, which is not so bad.


XGB Cross Validation RMSE
0.46
0.44
0.42
0.40
0.38
4
8
12
nround
class
test
train


XGB Cross Validation ERROR
0.1810
0.1805
0.1800
0.1795
0.1790
4
8
12
nround
class
test
train

6. Deployment

Our model could provide guidance to increase vaccination rates for Humana's members. After identifying est_age, we decided to make more campaigns targeting the old people. Many of them are hesitant to have vaccination because of the unavailability of getting the information about the vaccination, so the campaigns could be fully aware of the existence and importance of getting injected. As for the insight generated from 'rwjf_uninsured_adults_pct', we need to set different strategies for people under 65 with and without insurance. With regards to people who have more access to prescriptions for generic drugs, they may tend to have a more convenient way to combat with Covid instead of choosing the vaccination. We could educate them more about the safety and effectiveness of the vaccination and thus make it easier for them to take the vaccination.