Humana-Mays

Healthcare Analytics 2021 Case Competition

Team Cobras

Identifying Humana members most likely to be hesitant to the COVID vaccine

Table of Contents

1. Executive Summary

[1] The 2020 pandemic has taken a toll all around the globe, especially in the United States. Over 700 thousand people in the United States have died from the COVID-19 virus. Initially, some governments tried to increase awareness of the potential risks of the virus as well as the ways in which to combat it. Later on, the covid vaccine was developed and distributed all around the world. However, more than a year later, the pandemic is still as real as it was back when lockdowns were declared. [2] Scientists point out the fact that we, as a global community, have not yet reached the point of herd immunity. Herd immunity transpires when a large portion of a group of people becomes immune to a disease, making the spread of the disease less likely from person to person. At herd immunity, the entire community becomes shielded from the disease, not only the immune few. Reaching herd immunity is not an easy feat, given that a large portion of the population would have to be immune. What makes it even more difficult are the different variants of the virus that develop over time. Therefore, now, more than ever, it is critical for the population to get vaccinated. The pandemic will never end as long as there are a considerable amount of people without vaccination. Time is of the essence, as every day, thousands are still dying from the virus. [3] In developing countries, vaccine access has been an issue because of inefficient distribution, logistics, and overall lack of vaccines. [4] However, in countries such as the United States, the supply of vaccines is not an issue as they have enough supply for boosters. One might wonder why there are still eligible people who are unvaccinated.

For healthcare insurance providers such as Humana, it is essential to answer this question. Increasing COVID vaccination rates among its members is one of Humana's highest priorities. Humana is particularly focused on delivering vaccines to those underserved portions of the population. Nonetheless, there is a portion of the members of the population who are reluctant to receive the vaccine, either because of determinants regarding misinformation or lack of trust towards the government. The goal of this project is to assist Humana in identifying the members most likely to be hesitant to the COVID

vaccine. Our findings can help Humana design targeted outreaches for these hesitant members as well as prioritize to reach the most vulnerable and underserved populations to receive health solutions.

The first part of our project is delving into the data and understanding the problem at hand. Our first goal was to look at the dataset as a whole and see potential patterns and interesting characteristics of the different variables given. We had to make sense of a dataset containing hundreds of variables and tens of thousands of rows. We used exploratory data analysis in order to find patterns in the data and extract insightful observations. The data at hand was overwhelming given its breadth and depth. Therefore, we found it useful to categorize the different variables into 8 different categories. We then eliminated those variables which had more than 95% null values. Before starting our model building, we had to identify our target variable, COVID vaccination (1 if a member received the COVID vaccine, 0 otherwise). Moreover, we ran a correlation matrix of the variables for each category in order to further understand the relationship and possible connection among variables of each category as well as the correlation with the target variable. We then chose the variables most correlated with the target variable, for each category, and eliminated variables with low correlation to the target variable.

The second part of our project encompasses forming a model able to predict Humana members most likely to be hesitant to the COVID vaccine using health and socioeconomic data as input.
We analyzed the effectiveness of different machine learning models in order to find the best fit. In order to be consistent and be able to compare each model effectively, we used the AUC score as the key performance indicator. We started with three popular machine learning models (random forest, XGBoost, and LightGBM), and determined XGBoost to be the best model predictor.

2. Data Preparation

2.1.1 Data Understanding

To perform our analysis, we were given a training dataset of 974,842 rows and 368 columns.
In order to analyze the dataset at hand, we first used Google's BigQuery in order to handle the colossal magnitude of the dataset. We imported the dataset as a CSV file with all variables casted as strings. We then executed queries in order to understand the nature of the data. After BigQuery, we moved our dataset into Google Colab notebooks, where we would code most of our studies in the Python language. We used the Pandas and Numpy libraries for the exploratory data analysis stage of our project as well as seaborn for the visualization aspect of our findings. From there, we categorized the columns into seven main categories: medical claims, clinical condition, cms member data, credit data, demographic/consumer data, lab claims, and pharmacy claims.

2.1.2 Data Cleaning/Variable Reduction

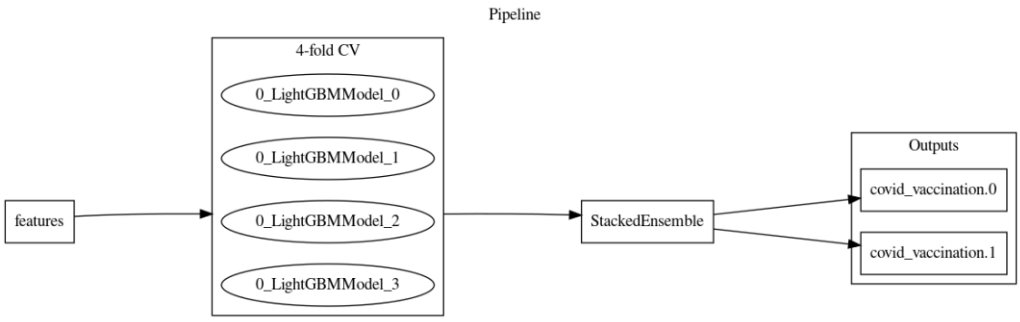
First of all, we dropped 109 columns in the dataset because most of them had above 90% zeroes. Columns that have a variance of zero do not influence the target feature of 'covid_vaccine'. Therefore, we removed the following feature columns. In each category, we ran a correlation matrix to find the correlation between the target variable and all other variables in that category. Then, we selected the top variables in each category in order to clean the data. After that, we imputed the values of the missing features 'Na'. For features with a high number of outliers, we replaced their missing values with the median value in those features. Finally, for selecting the features we created a correlation matrix and a heatmap and selected the top 30 features as our potential parameters for machine learning models.

3. Modelling

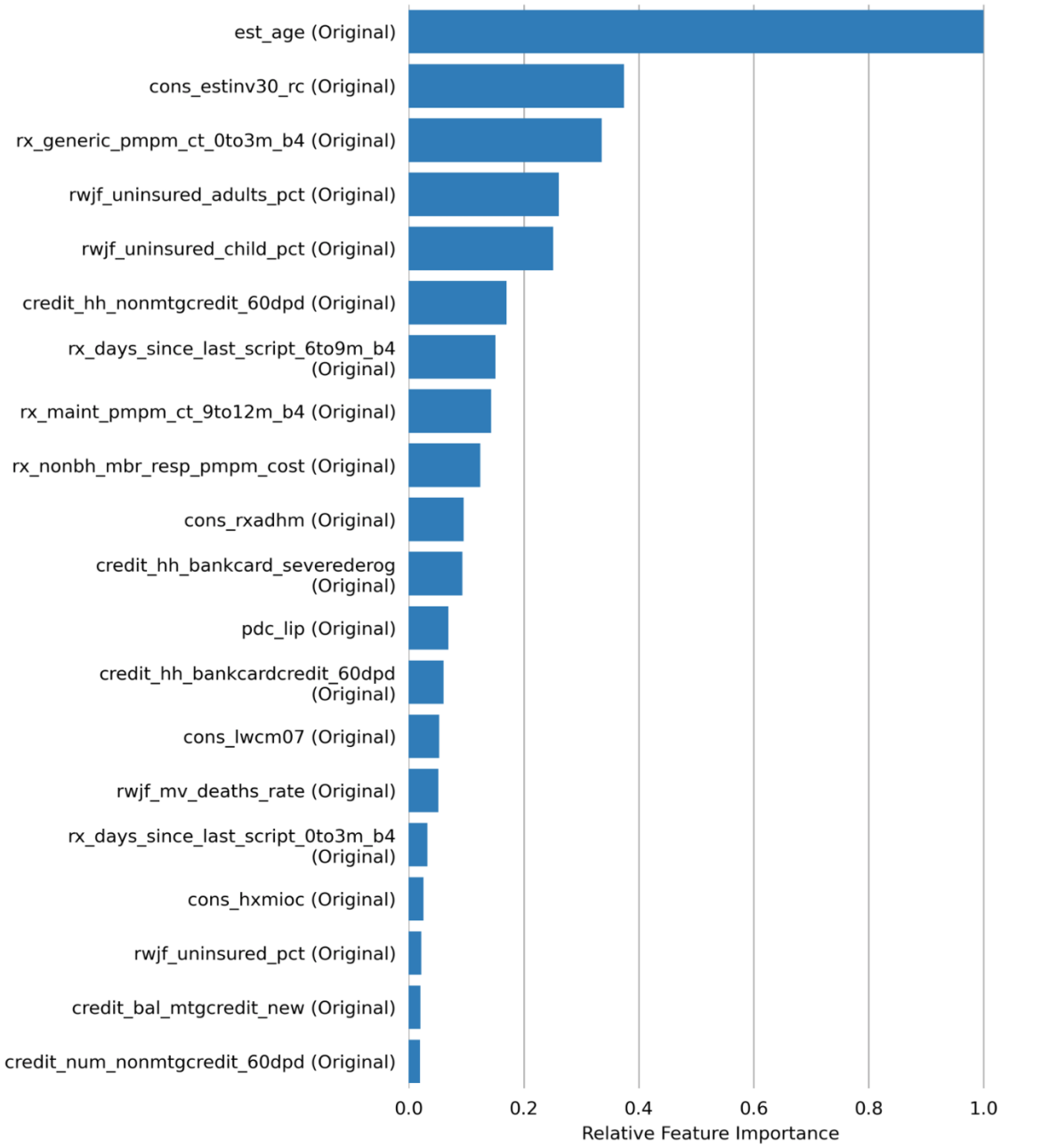
3.1 Model Selection

Before we tried different models and hyperparameter sets, we took reference from H2O.ai, the automated machine learning module, to fit our dataset with selected features. During the Model and Feature Tuning Stage, Driverless AI evaluates the effects of different types of algorithms, algorithm parameters, and features. The goal of the Model and Feature Tuning is to determine the best algorithm and parameters to use during the Feature Evolution Stage. The module uses a genetic algorithm to find the best set of model parameters and feature transformations to be used in the final model: a bagged ensemble (pasting) of 1 LightGBM Model across 4 folds.

Final StackedEnsemble pipeline with ensemble_level equal to one transforms 23 original features into 26 features. In each of two models, each fits on four internal holdout splits then linearly blended:

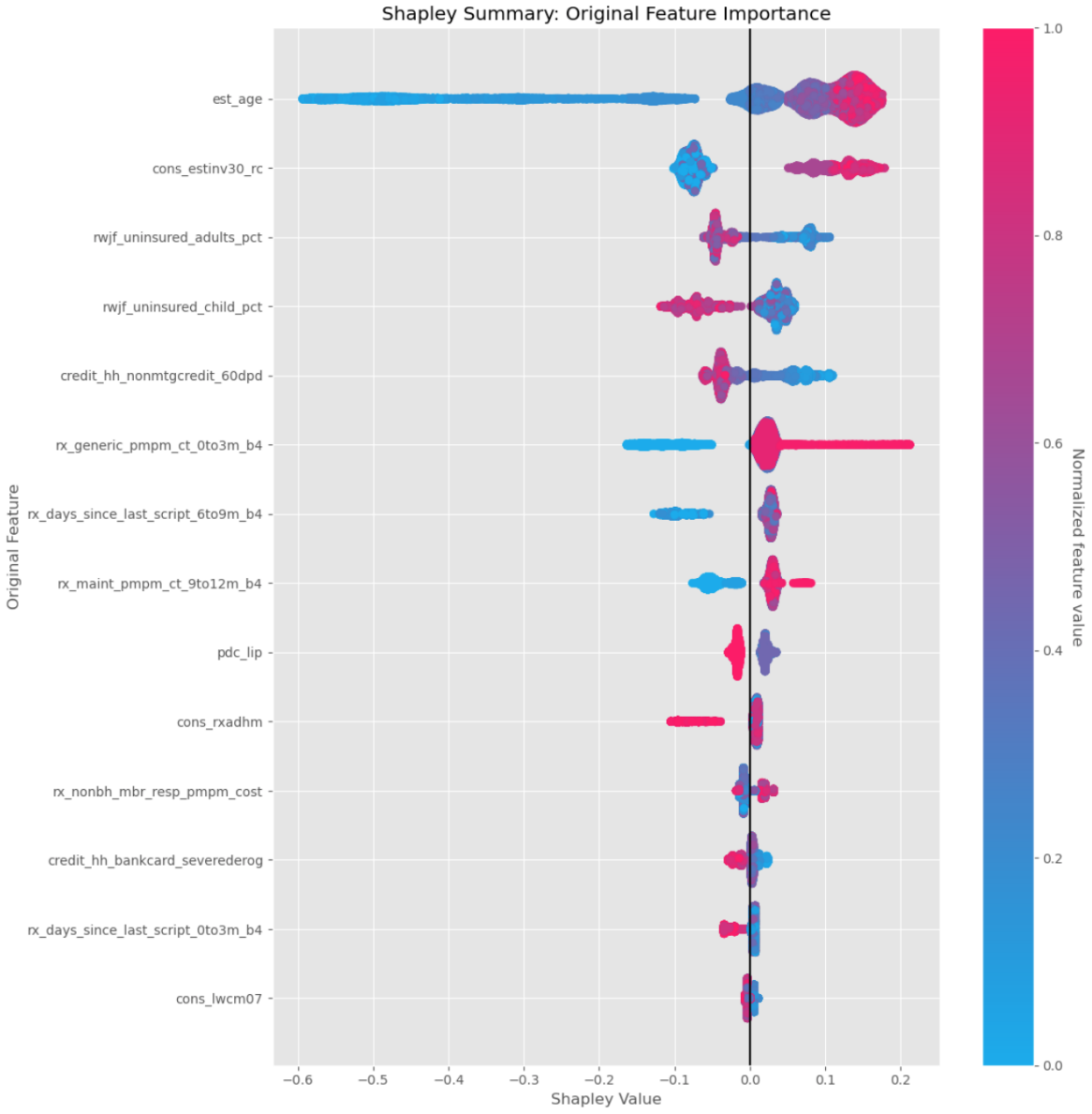


The overall feature importance in this model is shown below:



After identifying the top performance indicators, it is time to understand the relationship between those variables with covid_vaccine. Therefore, in the Shapley Values Plot, a technique with credible theoretical support presents consistent global and local feature contributions. For regression problems, local Shapley feature contributions plus the bias term sum to the final model's prediction. For classification problems,

they sum to the prediction before applying the link function. The following Shapley summary plot is created from a random sample of 10000 rows (the autodoc_pd_max_rows configuration controls random sampling for this plot):



As we can see, variable est_age, cons_esyinv30_rc and rx_generic_pmpm_ct_0to3m_b4 influence the covid_vaccine positively. This means the probability that a patient taking a covid vaccine shot will increase when these variables increase.

To further select and tune the best model, we examined three different binary classification algorithms: Random Forest, XGBoosting, and Light GBM. Then we evaluated each model using AUC scores while we kept the training size and the test size the same for each model. One of the machine learning models that we selected was Random Forest. We imported RandomForestClassifier from the sklearn package. Not only was Random Forest useful in the selection of the most important features, but also it was another model that could be a good predictor of the target variable. The model creates an ensemble of multiple decision trees for different kinds of classifying objects. While each decision tree may run the risk of overfitting the data, averaging a large number of trees provides an unbiased and lower variance solution, as well as a better predictor than a singular tree. As mentioned earlier, Random Forest also provides a ranking on the variable importance out of the variables selected. In our case, after running a correlation matrix and performing variable reduction, we provided the model with 30 variables, out of the original dataset. For this model, the parameters that we used were criterion = "entropy" and class_weight = "balanced". However, Random Forest did not prove to be the optimal model as it resulted in a poor 0.5 test AUC.

Another one of the models was XGBoosting. XGBoost implements the gradient boosting decision tree algorithm, which improves the execution speed and model performance. XGBoost requires parameter tuning to improve and fully leverage its advantages and we will discuss it in the Model Tuning section. Finally, we evaluated LightGBM. LightGBM is a gradient boosting algorithm that uses tree-based learning algorithms with many advantages, such as faster training speed and higher efficiency, better accuracy, and capability of larger-scale data.

Random Forest had the lowest AUC score while XGBoosting and LightGBM had very close AUC scores, therefore, we decided to continue with XGBoosting and LightGBM in the model tuning phase.

3.2 Model Tuning

XGBoost and LightGBM were tested with a baseline set of parameters before the hyperparameter tuning.

Model	Hyperparameter Tuning	ROC-AUC Score
XGBoost	subsamples= 0.23, scale_pos_weight= 6, n_estimators=130, max_depth= 7, learning_rate= 0.1, colsample_bytrees=0.8, colsample_bylevel= 0.1	0.660189
LightGBM	learning_rate=0.11,max_depth= 10,random_state=42,metric='auc ',subsample=0.7,max_leaves=4,t ree_method='gpu_hist',colsampl e_bytree=0.9	0.659994

Compared to LightGBM, XGBoost has a slightly higher ROC-AUC value, and thus it is proved to be a better predictive model for analysis than LightGBM on our dataset. We also used XGBoost to estimate the feature importance.

3.3 Final Model

XGBoost is identified as the optimal model to predict the holdout dataset. From the Figure 1 below, we obtained an AUC score of 0.6602 which was tested on the validation set.

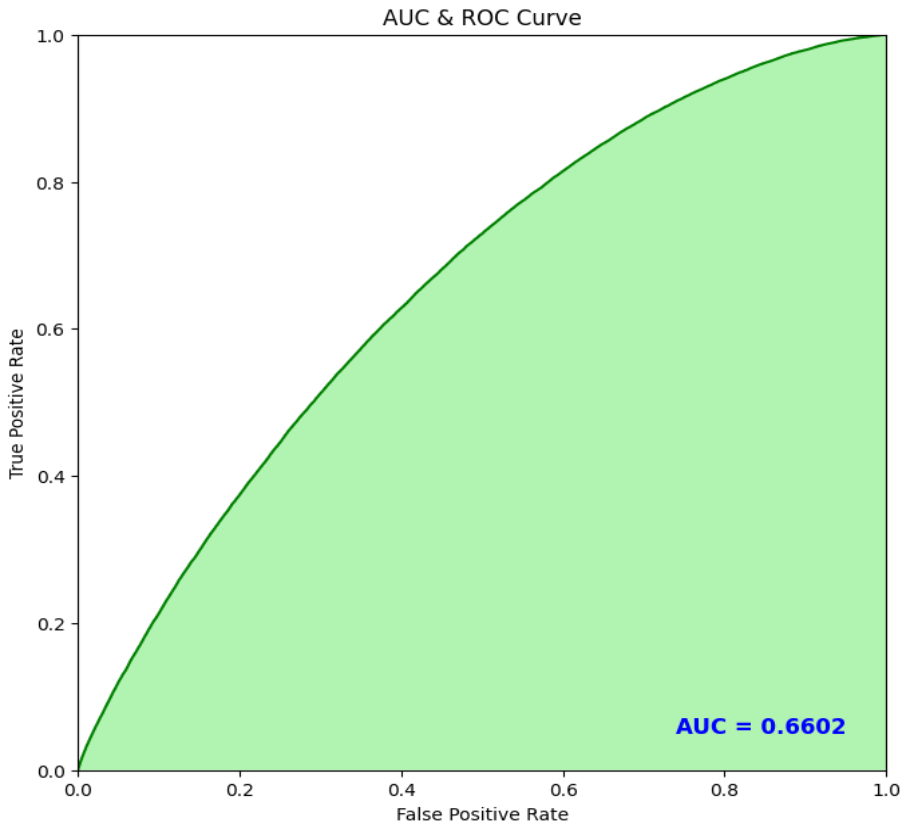
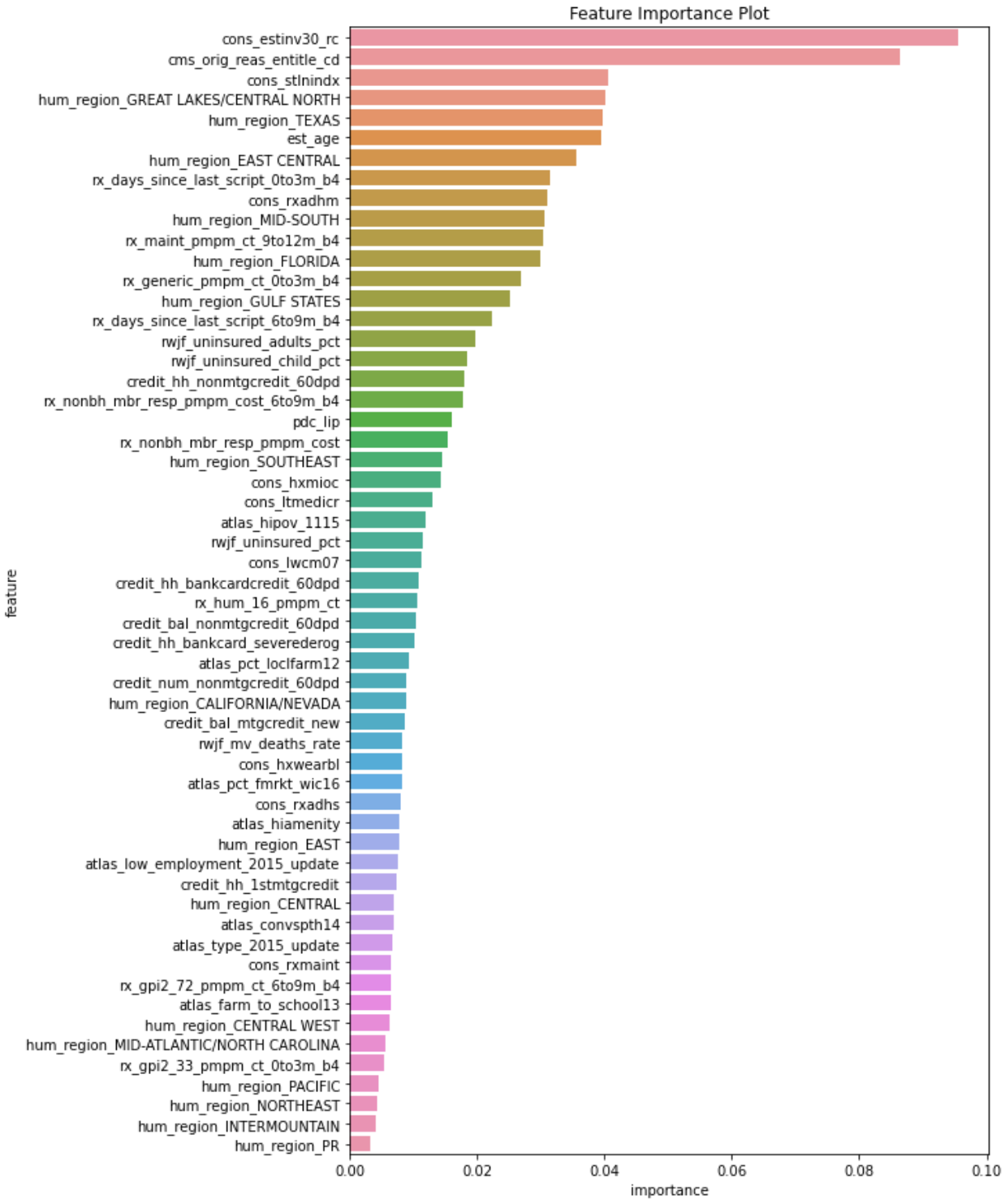


Figure 1: ROC Curve for the Tuned XGBoost Model

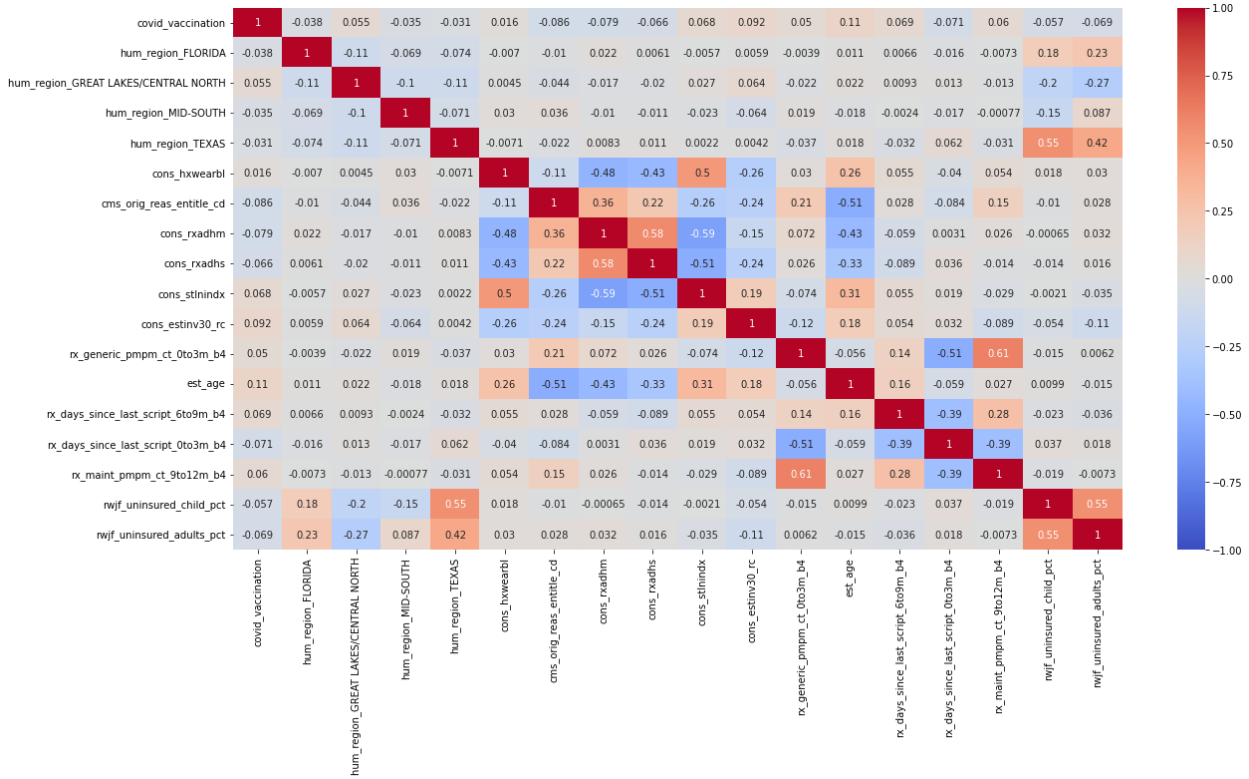
3.4 Feature Importance

In order to better grasp the behavior of the Humana members who are more hesitant to receive the vaccine, we had to look at the most influential features in our model. Using XGBoost, we were able to rank the features with the highest impact on the target variable, covid vaccination.



We find it useful to group the above features in order to make our analysis more interpretable and readable. We divided the top features into categories: health, location, demographic.

3.4.1 Top Features: Analysis



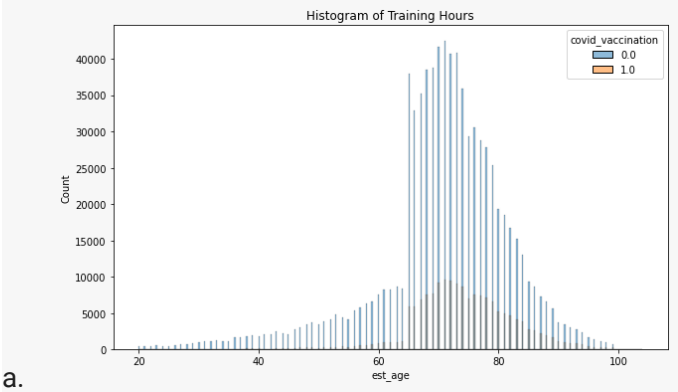
We used a correlation matrix in order to compare the relationship, either positive or negative, of each top feature against the target variables. Here, we analyzed the top 17 features.

DEMOGRAPHIC

Estimated Household Investable Assets Recorded ranges from 2500 to 1000000 with 10 tiers. The higher the number, the more household investable assets the individual has. It is the most important factor when it comes to explaining an individual's willingness to get vaccinated. Individuals with more household investable assets are more willing to get vaccinated, on the other hand, individuals who have less household investable assets are more likely hesitant to get covid vaccination. Therefore, individuals with fewer household investable assets would be one of the characteristics of the target population when insurance companies like Humana promote the covid vaccine.

As for the code indicating the original reason for entry into Medicare (cms_orig_reas_entitle_cd), the correlation between this variable and covid-vaccination is -0.085884. If an individual enters Medicare because of disability or end-stage disease, they are less likely to get a vaccine. [5] According to the research by Kidney Organization, it shows that people who are on immunosuppression medications for the treatment of advanced kidney disease and kidney transplant recipients, may not receive the same level of protection, also known as antibody immunity, from the COVID-19 vaccine as people who are not on immunosuppressive medication. Therefore, people with End Renal disease will hesitate to get the vaccination.

From the histogram plot, we observe that the age of Humana members mostly ranges from 60 to 90. Individuals in the senior group (60-90-year-old) are more likely to get the vaccine. This can be explained by the higher risk for seniors when exposed to the virus.



a.

Student Loan ranges from 0 to 9 with 9 being the category that has the most student loans and 0 being the category that has the least amount of student loans. During our feature analysis, we found that people's willingness to get vaccinated has a positive correlation of 0.067751 with the number of their student loans. That means people with more student loans will be more likely to get vaccinations versus people with no student loans. We suppose that people with student loans most likely are those who received high education and thus are more open to vaccines. Therefore, individuals with less or no student loans could be a characteristic of the vaccine promotion target group.

LOCATION

Given the importance of feature analysis, the location features with the most impact on the target variable were: Great Lakes/Central North, Texas, East Central, Florida, Mid-South, and the Gulf States. These are the geographic information for the Humana members. By performing a correlation analysis we came to the conclusion that different locations have different behaviors in regards to hesitancy to the vaccine. The Southern and Mid-Southern states had a similar negative correlation while the Northern States have a positive correlation with getting vaccinated. [6] Based on a Fortune.com article, some key reasons for vaccine hesitancy in the states mentioned above are worries about side effects and anxieties regarding the fast pace of the clinical trials. One important element to point out is that vaccine opposition is more prevalent in young adults and in states where the population majority is Republican.

HEALTH

The feature regarding days since the last prescription in the past three months prior to the score date has a negative correlation with the covid vaccination status, which is contradictory to the following feature. Intuitively, it should be a positive correlation because people who are taking prescriptions should be more cautious about their health, hence they should be more likely to get a vaccine, which means it is a positive correlation. However, the model tells us that it is a negative correlation. [7] In fact, it is because people who got vaccinated in January 2021 were advised by doctors not to take certain prescriptions because there were concerns that prescriptions might reduce the efficacy of the vaccine. As a result, their correlation is negative.

The feature regarding days since last prescription in the past sixth to ninth month prior to the score date has a 0.0693 correlation with covid_vaccination status. This correlation means people who have been taking prescriptions are more likely to get vaccinated because they are more concerned about their health. One of the health-related variables was the count per month of prescriptions related to maintenance drugs in the past ninth to the twelfth month prior to the score date. The relationship between this variable and the target variable was positive. This relationship makes sense because the individuals who take maintenance drugs usually have a chronic disease. [8] Since the data here is six to nine months prior to

March 2021, it means that those individuals with health conditions would be more likely to be in favor of the vaccine as they are more vulnerable to the virus. Similarly, the feature regarding adherence to maintenance medication displayed a positive correlation with the target variable. The reasoning is analogous to the feature regarding maintenance drugs prescriptions. Patients with chronic conditions are more likely to favor the vaccine as the coronavirus poses a greater health risk for older adults, especially those with complicating preexisting conditions and weaker immune systems.

Both the uninsured adults and uninsured children have a negative relationship with the covid_vaccination. The larger the portion of adults (age<65) or children (age<19) without health insurance is, the less likely they are to get the vaccination. [9] Even though the government has provided funding to reimburse medical providers for the cost of administering the COVID-19 vaccine to uninsured adults/children, some providers might opt to bill the patient directly instead. Therefore, there are concerns that could make the uninsured individuals hesitant to seek vaccination if there is any chance they have to pay.

4. Actionable Insights and Recommendations

In this section, we will provide our recommendations as to increase vaccination rate based on our analysis for Humana's Medicare members.

4.1 Proposed Solutions

- As for members with less or no investable assets, they will be most incentivized to get the covid-19 vaccine if Humana can provide extra health care benefits covered by the insurance plan for those who get the vaccination, such as lower copay, more free care, Rx discount, and free telehealth just to name a few.
- Per our analysis, members in the southern and mid-southern states are relatively hesitant to get a Covid-19 vaccine. With these geo-locations of targeted groups, local governments or Insurance companies like Humana can put advertising on social media directly targeting people in these regions to help them better understand the facts about the covid-19 vaccine and quash misinformation around the covid-19 vaccine.
- Families with uninsured children or adults are less likely to get a covid vaccine because they are afraid that they would get a bill from insurers. Some measures can be taken:
    - Therefore, one way to increase people's willingness to get a shot is to increase awareness among members that the covid vaccine is free.
    - At the same time, insurers such as Humana should make sure that no member is billed for getting the covid vaccine. As a result, people will not be discouraged from seeking the covid vaccine.
    - [10]Also, medical providers may build trust with patients and share information with patients in a more effective way, such as holding virtual 'office hours,' and leveraging digital health tools (Mckinsey).

    - Physicians may also consider building upon vaccine services (for example, being able to schedule appointments directly with the doctor, getting follow-ups post-vaccination) (Mckinsey)
- Since most Humana members are seniors (age>65) and they have higher health risks when exposed to the virus, Humana can first pay more attention to the senior group on getting vaccinations. Older people may be hesitant to get vaccines due to a lack of expertise with modern technology causing them to skip the vaccine. It is suggested that Humana educate the seniors in making online appointments for vaccines. Through getting contact with the seniors, Humana at the same time can outreach their family members to increase their awareness of getting vaccinated and help them with the vaccination process.

5. Conclusion

In predicting what type of Humana's Medicare members were most likely to be hesitant to get a covid vaccine, we analyzed 367 features and filtered out 311 features based on the completeness of the data, correlation strength, feature importance as well as business acumen. Ultimately, 56 features were selected and used to build machine learning models. Three machine learning models were created and evaluated. The XGBoost model was selected with an AUC score of 0.660189. Our research and analysis indicate that people with certain characteristics will be less likely to get the covid vaccine, and we think Humana should put more effort and resources into these types of members to help them understand the necessity and importance of the covid vaccine.

6. References

[1]https://www.worldometers.info/coronavirus/country/us/

[2]https://www.bloomberg.com/news/articles/2021-08-13/the-world-may-never-reach-herd-immunity-against-covid-19

[3]https://www.unaids.org/en/resources/presscentre/featurestories/2021/march/20210310_covid19-vaccines

[4]https://apnews.com/article/coronavirus-pandemic-vaccine-boosters-children-demand-us-ca6d151522986900211f467740211994

[5]https://www.kidney.org/coronavirus/vaccines-kidney-disease

[6]https://fortune.com/2021/09/02/covid-vaccine-by-state-hesitancyrankings-delta-variant/

[7]https://www.healthline.com/health-news/these-prescription-drugs-may-reduce-efficacy-of-covid-19-vaccines#What-can-we-do-mitigate-this-problem

[8]https://www.pewresearch.org/science/2020/12/03/intent-to-get-a-covid-19-vaccine-rises-to-60-as-confidence-in-research-and-development-process-increases/

[9]https://www.verywellhealth.com/covid-19-vaccine-for-uninsured-5090206

[10]https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/whos-left-engaging-the-remaining-hesitant-consumers-on-covid-19-vaccine-adoption