HUMANA-MAYS HEALTHCARE ANALYTICS 2021 CASE COMPETITION
DETERMINING WHICH MEMBERS ARE LIKELY TO REFUSE COVID VACCINATION

CAPSTONE PROJECT

DECEMBER 17, 2021

SUBMITTED TO: PROF. AMIN GHARIPOUR
(DEPARTMENT OF MANAGEMENT AND INFORMATION SYSTEMS)

BY: KUNAL SHARMA

Table of Content

1. Executive Summary

This study seeks to assist Humana in identifying members who are more likely to be hesitant to the COVID vaccine, gaining insights from the data, and, most importantly, providing advice and viable solutions to encourage vaccination among the hesitant members' sub-segments. Furthermore, in order to establish justice and equity, we attempt to eliminate possible bias inherent in the data. To begin, we determined the target variable "covid_vaccination" ("vacc" if vaccinated, and "no vacc" if not vaccinated). The target variable was severely skewed. The traits were then separated into eight groups and analyzed using EDAs and statistical tests. The data was then preprocessed to remove unnecessary columns, transfer variable data types, and deal with missing values, allowing the model to capture underlying trends. Following data preparation, we considered the benefits and drawbacks of many models before selecting the best one for our study.

The reluctant members were then separated into four sub-segments, and specific recommendations were made appropriately. Humana may use social media platforms like as Twitter and TikTok to disseminate scientific vaccination information to young members. Humana may cooperate with local clinics and pharmacies to give monetary incentives to promote vaccinations to members in specified locations 2 where the vaccination rate is low. For members who are unsure about the safety of vaccines, emails/brochures can be sent out to educate the audience about the safety and effectiveness of immunizations. Humana can provide certain 24/7 walk-in services to meet the requirements of low-income people who have limited access to immunizations.

2. Case Background

Coronavirus disease (COVID-19) is a viral infection caused by the SARS-CoV-2 virus. It first appeared in the United States in early 2020 and quickly spread to become a continuing epidemic, with 42,966,938 verified cases by the end of September 2021. (Elflein, 2021). The enormous epidemic has wreaked havoc on society and the economy. As the pandemic spreads, it is becoming increasingly critical to boost vaccination rates in our community. In the United States, approximately 56.5 percent of the population is completely vaccinated, leaving half of the population unprotected and vulnerable to the virus (Ritchie et al., 2020). Vaccinating a large population successfully helps to limit the spread of the virus and boost community immunity. Humana, as a significant healthcare business, is likewise concerned about COVID-19 immunization among its members. Even though Humana makes every attempt to provide immunization chances to the most vulnerable and disadvantaged communities, some members are still unwilling to provide vaccine. Many causes might account for the reluctance, including a lack of faith in science and insufficient knowledge about the benefits and drawbacks of vaccination. As a result, this research was created to investigate the factors that influence one's vaccination attitude. We intend to identify members who are most likely to be resistant to the COVID vaccination and offer methods for Humana to explain the benefits of immunization to various demographics.

3. Data Preparation

3.1 Data Understanding

This Project seeks to assist Humana in identifying members who are more likely to be hesitant to the COVID vaccine, gaining insights from the data, and, most importantly, providing advice and viable solutions to encourage vaccination among the hesitant members' sub-segments.

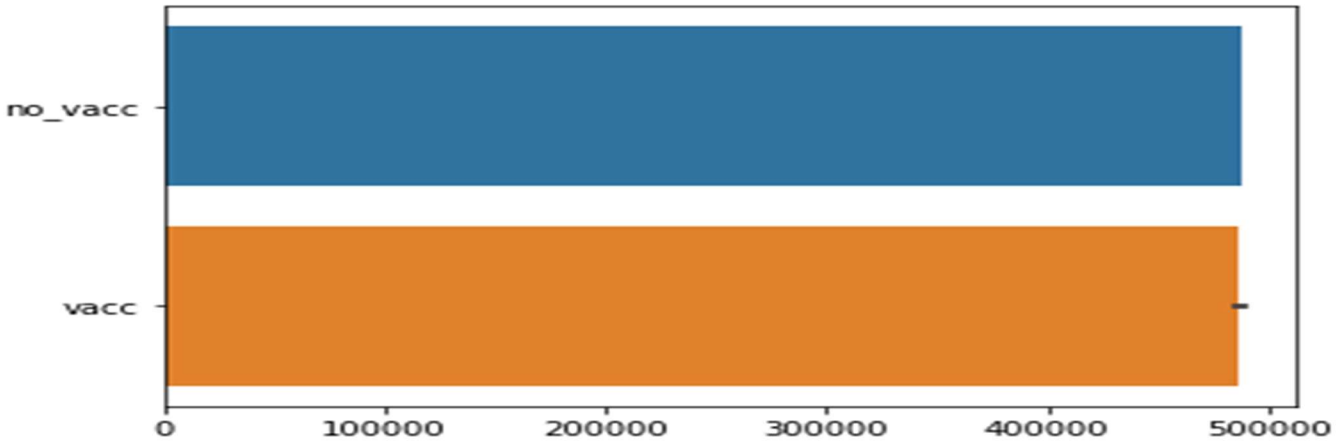
Furthermore, we In order to establish justice and equity, we seek to reduce any biases in the data.

To conduct our research, we were provided a training dataset of 974,842 rows and 368 columns. The data was collected at the customer level and included eight distinct types of characteristics, including Medical Claims, Pharmacy Claims, Lab Claims, Demographics, Credit Data, Condition Related Features, CMS, and Others. Following some early analysis of the data, we discovered the following insights:

- The data was extremely skewed with the target variable in the ratio of 805,389 for class 0, persons who are hesitant to be vaccinated, and 169,453 for class 1, those who are not hesitant to get vaccinated.
- Missing values were represented in two ways: "*" and NAN. After examining the distribution for both forms of missing values, we discovered that in columns where "*" and NAN coexisted, "*" accounted for a relatively tiny part of the total (less than 1 %).

To begin, we determined the target variable "covid vaccination" ("vacc" if vaccinated, and "no vacc" if not vaccinated). The target variable was severely skewed. The traits were then separated into eight groups and analyzed using EDAs and statistical tests.

Below graph show distribution of the Data Set w.r.t target Variable.



The data was then preprocessed to remove unnecessary columns, transfer variable data types, and deal with missing values, allowing the model to capture underlying trends. Following data preparation, we considered the benefits and drawbacks of many models before selecting the best one - LightGBM - to carry out our research.

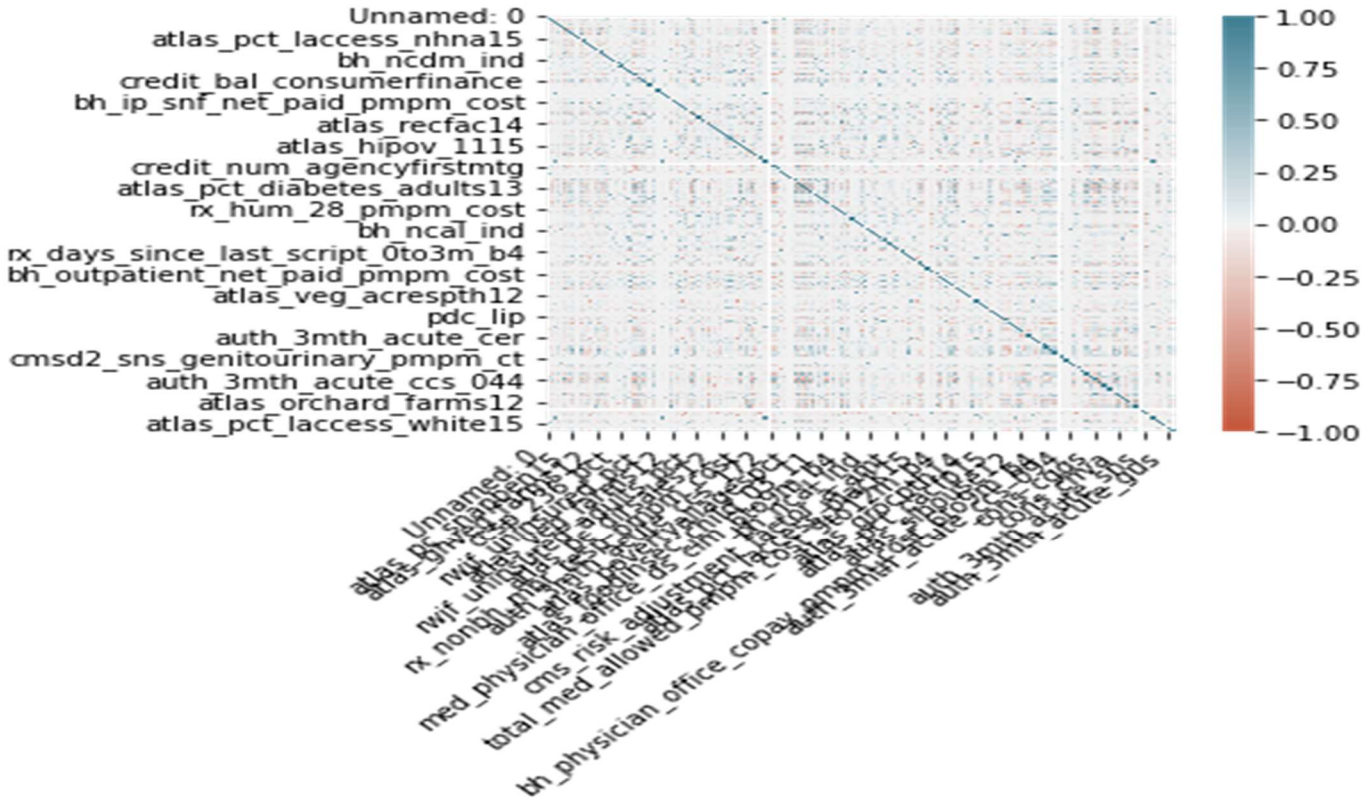
3.2 Feature Engineering

After doing feature importance analysis, we identified some surprising trends. Member age, geographic location, (Risk Adjustment Factor) RAF quantity, and proportion of persons under 65 without health insurance, among other factors, all have a major influence on whether a member is hesitant to obtain the immunization.

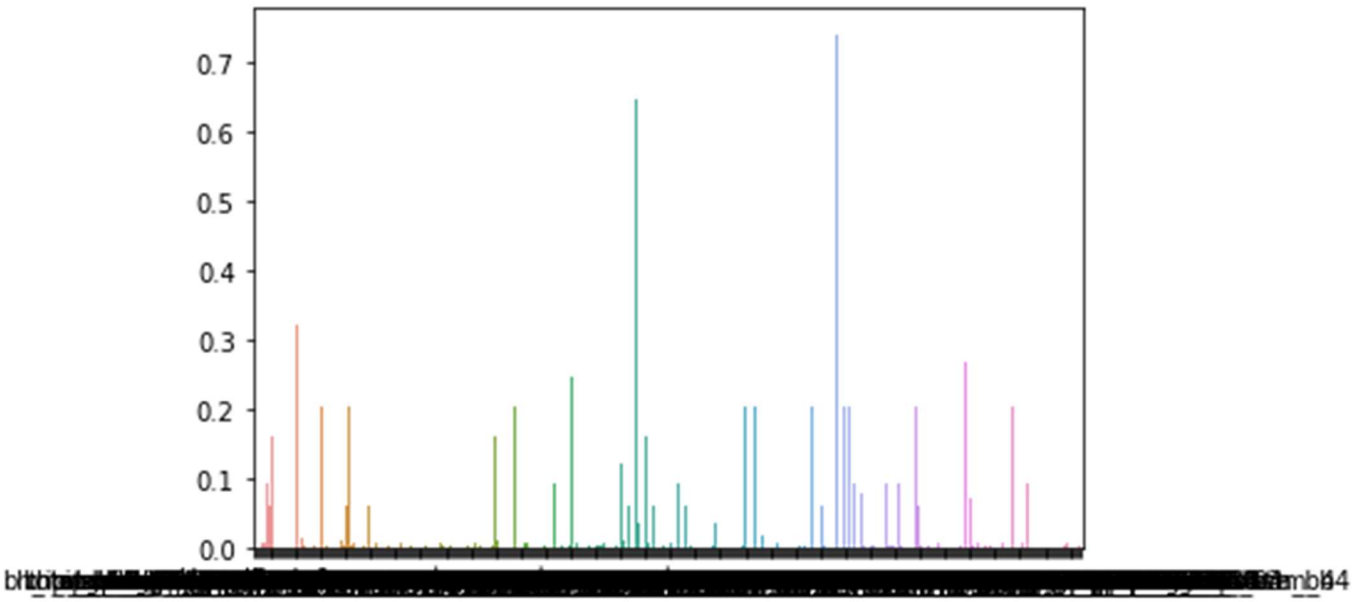
Many columns had a combination of data types. Some were numerical but arrived in the shape of objects, while others were categorical but came in the form of numbers. We converted all of these columns to the appropriate data types. We replaced all "*" with NAN for the two types of missing values "*" and NAN.

We only maintained features that had more than one unique value during the feature selection process. Furthermore, if a column had just two unique values and the frequency of one value was greater than 99.9%, we deleted the column as well. Features with zero or very minor variance will have no effect on our desired feature of "covid vaccination." As a result, we eliminated the following 40 columns.

Below is the correlation graph b/w the features.



In Below fig. each bar size represents the high number of Null Values for each feature.



We classified the features as numeric and category.

3.2.1 For Numeric features:

To save memory and speed up processing, we changed various data types to "Int64." We left the missing values alone because our model tolerated null values. We also standardized all of the numerical characteristics for improved performance.

3.2.2 For Categorical features:

I have imputed the most frequent category to columns with 10% or more missing values, and we created a new category "blank" for those with fewer than 10% missing data. Then, for two reasons, we chose Label Encoding for categorical features. One reason for this is that encoding more than 50 categorical features with high cardinality using approaches such as one-hot encoding or dummy variable encoding may result in a huge sparse matrix, which may have a detrimental effect on model performance.

Further, We only maintained features that had more than one unique value during the feature selection process. Furthermore, if a column had just two unique values and the frequency of one value was greater than 99.9%, we deleted the column as well. Features with zero or very minor variance will have no effect on our desired feature of "covid vaccination." As a result, we eliminated the following 40 columns.

auth_3mth_post_acute_rsk	auth_3mth_bh_acute_men	auth_3mth_post_acute_ben	auth_3mth_acute_hdz	
auth_3mth_acute_ccs_048	auth_3mth_acute_men	auth_3mth_acute_end	auth_3mth_rehab	auth_3mth_hospice
auth_3mth_acute_ccs_086	auth_3mth_dc_hospice	auth_3mth_acute_cer	auth_3mth_acute_ccs_030	
auth_3mth_acute_dia	auth_3mth_acute_skn	auth_3mth_acute_ccs_067	auth_3mth_acute_neo	
auth_3mth_acute_ccs_043	auth_3mth_post_acute_vco	auth_3mth_acute_cir	auth_3mth_post_acute_dig	
auth_3mth_acute_ccs_094	auth_3mth_post_acute_hdz	auth_3mth_post_acute_cad	auth_3mth_acute_ccs_172	
auth_3mth_acute_ccs_044	auth_3mth_acute_ccs_154	auth_3mth_post_acute_ckd	auth_3mth_post_acute_res	
auth_3mth_post_acute_ner	auth_3mth_acute_inf	auth_3mth_acute_ccs_042	auth_3mth_acute_cad	
auth_3mth_post_acute_inf	auth_3mth_post_acute_cir	auth_3mth_acute_sns	auth_3mth_acute_inj	
auth_3mth_post_acute_end	auth_3mth_acute_ccs_153	auth_3mth_acute_gus		

4. Modelling

We tried the following 3 models on our dataset:

1. Random Forest Classifier

2.Cat Boost Classifier

3.LightGBM

We trained all of our models on a 75-25 train-test split with 4-fold cross-validation for our dataset. Because our topic is a binary classification problem, we employed the accuracy, ROC-AOC, and log-loss metrics to evaluate our model. Our main goal was to reduce the quantity of false positives.

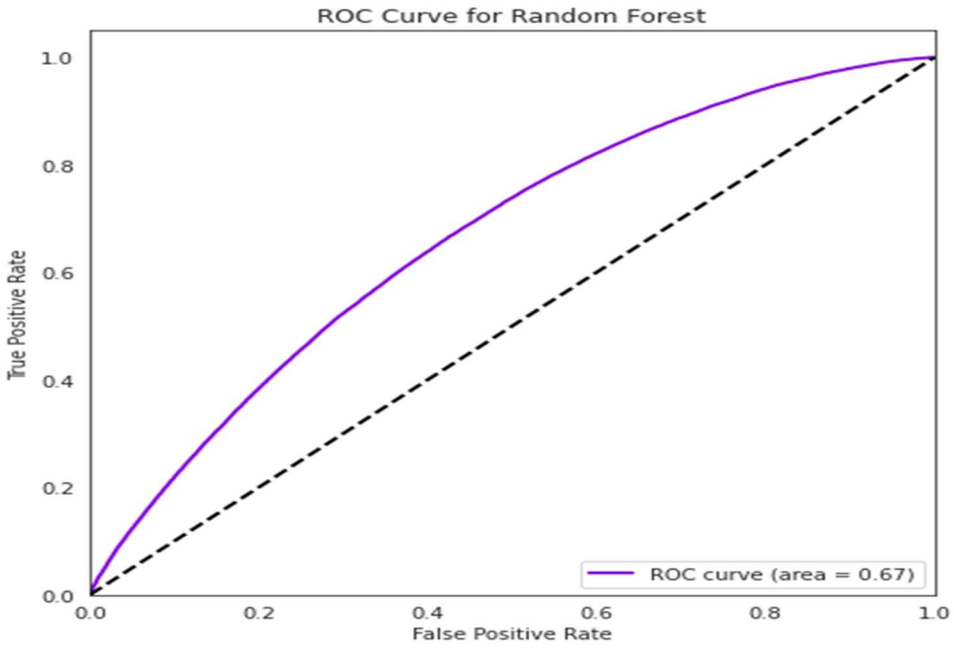
4.1 Random Forest Classifier

The decision tree is the fundamental unit of random forest classifiers. The decision tree is a hierarchical structure constructed from the characteristics (or independent variables) of a data collection. The decision tree is divided into nodes based on a measure connected with a subset of the characteristics. The random forest is a set of decision trees connected with a set of bootstrap samples created from the original data set. The nodes are divided according to the entropy (or Gini index) of a subset of the characteristics. The subsets formed by bootstrapping from the original data set have the same size as the original data set. Breiman's articles include extensive information on random forest classifiers (Breiman, 1996, 2001).

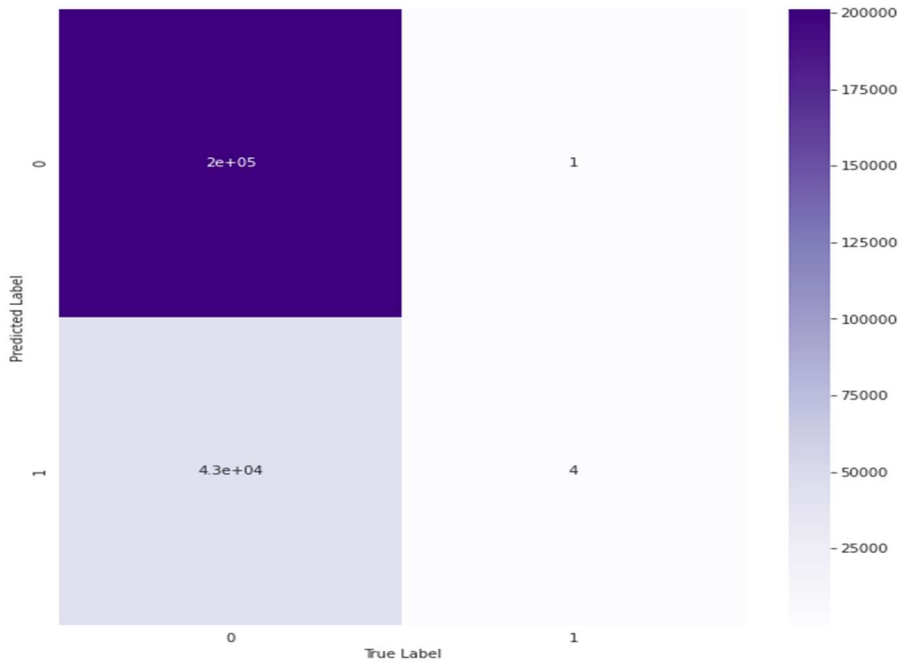
The bootstrapping strategy, as indicated by Suthaharan, aids in the building of a random forest with a given number of needed decision trees in order to increase classification accuracy through the notion of overlap thinning in the normal random forest approach (2015). The best trees are then chosen using a voting process using a technique known as bagging (bootstrap aggregate). The suggested cognitive computing architecture uses the traditional random forest technique.

Below is the Calculated important result of these models on the given train data.

AUC: 0.67



ACCURACY: 0.8251207372666807



4.2 Cat Boost Classifier

Python, being a multi-purpose programming language, supplies us with a variety of functions and modules that we may use to construct and obtain our data in a customized way.

When it comes to data science and machine learning, Python has a number of modules that instill the behavior of the machine learning algorithm and provide us with accurate results. Aside from machine learning algorithms, it also provides us with a variety of ways for preparing data for modeling and visualization.

We deal with regression as well as categorical data values in Machine Learning, i.e. numeric and categorical values. When it comes to categorical data, we frequently need to process them in order to get them in a numeric format so that the values may be categorized. Because of the vast data values that fluctuate according to the dataset, this work might be laborious at times.

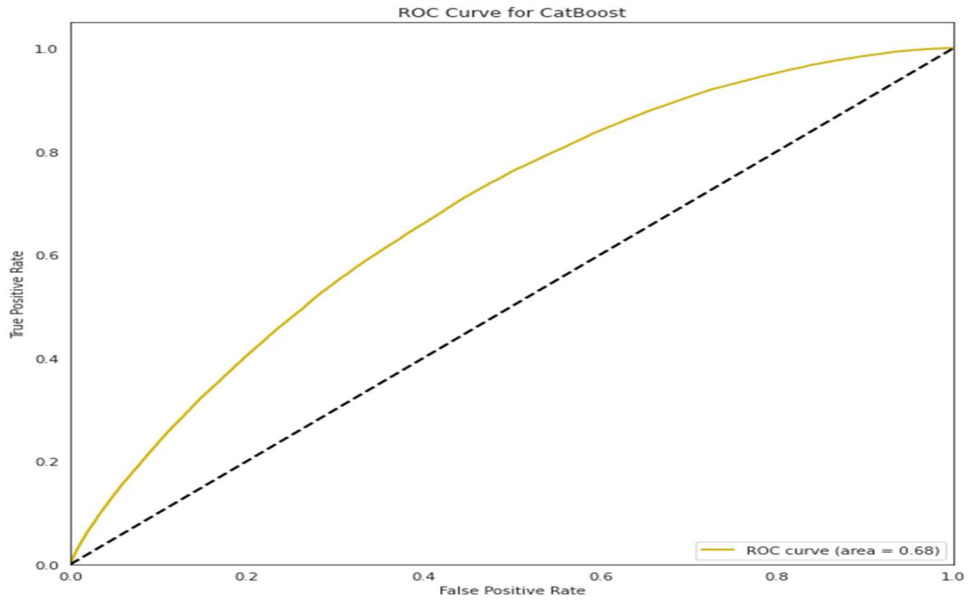
In the context of this observed difficulty, we shall introduce the Python Catboost module.

Catboost Model is a strong, scalable, and durable machine learning model that allows us to improve performance by combining the gradient boosting method with decision trees. It is also accessible for both categorical and continuous data values.

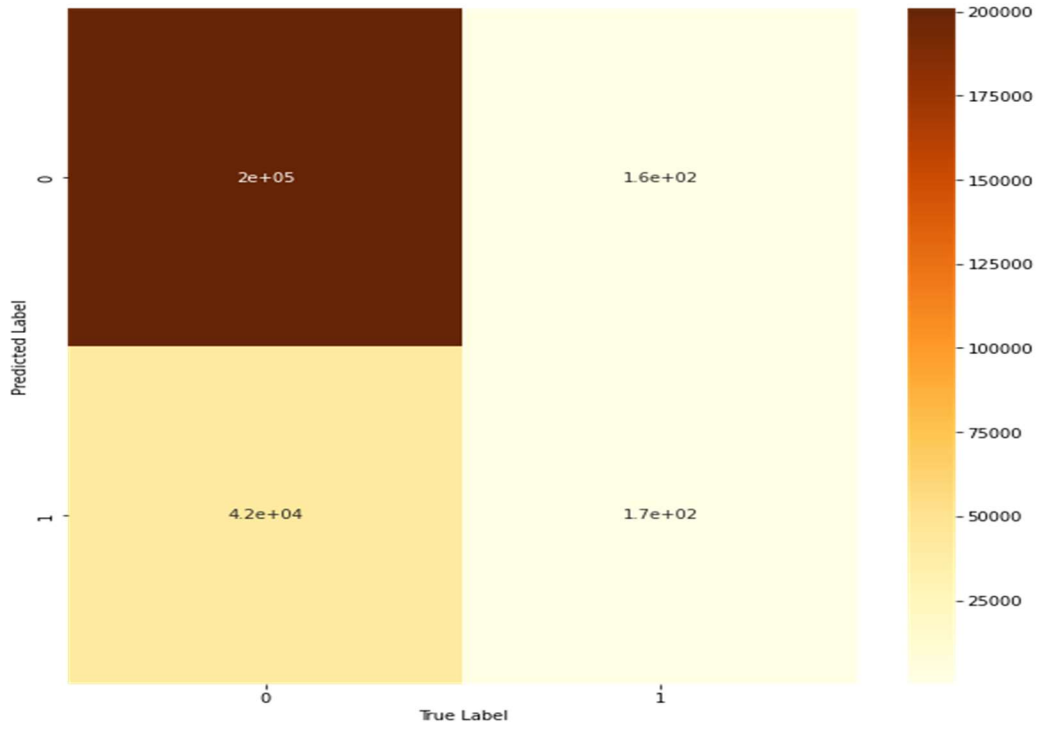
When we dive into categorical values, Catboost Classifier minimizes our cost of data translation from categorical data type to numeric form and also starts the model development process. It automatically enables and manages categorical characteristics or variables, as well as treats them.

Below is the Calculated important result of these models on the given train data.

AUC: 0.6819



ACCURACY: 0.8251617694728592

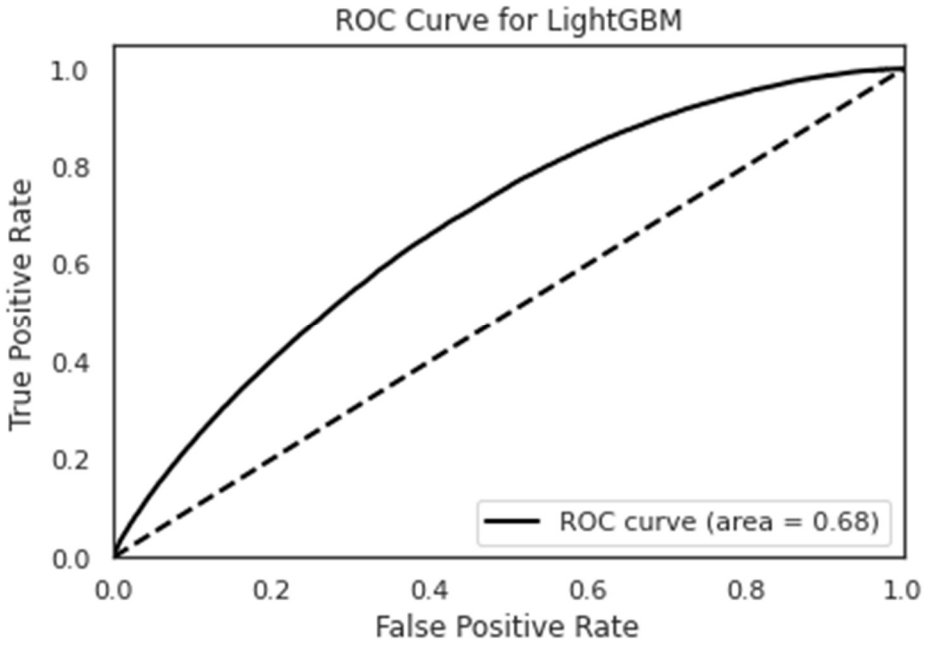


4.3 LightGBM

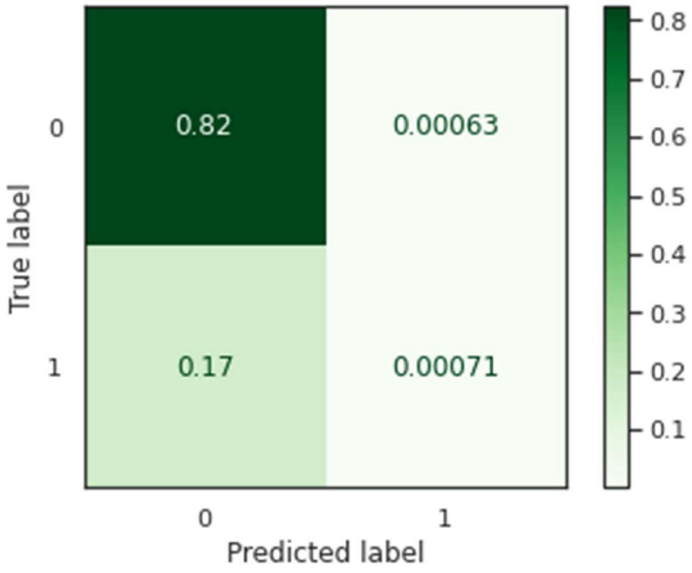
LightGBM is a framework for gradient boosting that use tree-based learning techniques. It is regarded as a fast-processing algorithm with very good accuracy. Furthermore, as previously stated, LightGBM handles missing data by default, allowing null values in our data. LightGBM also provides acceptable accuracy with integer-encoded categorical features, allowing us to encode our categorical features using Ordinal Encoding (Advanced Topic - LightGBM, 2021). We picked the LightGBM model for our predictive analysis because of all the above benefits and ease.

Below is the Calculated important result of these models on the given train data.

AUC:0.68



Accuracy: 0.8251404920171843

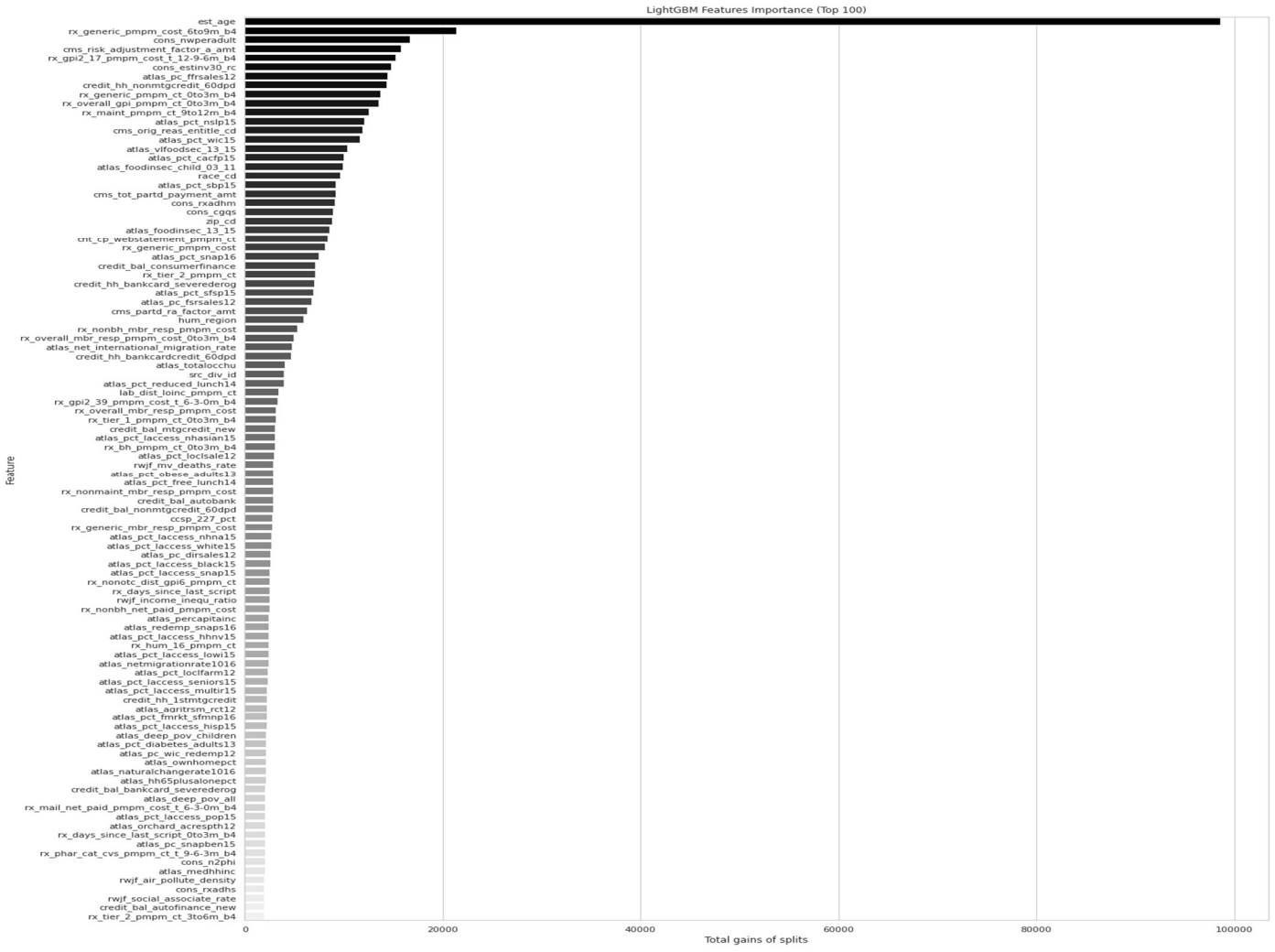


Evaluation Metric Chart for Models

	LightGBM Classifier	Random Forest Classifier	CatBoost Classifier
Accuracy	0.8251404920171843	0.8251207372666807	0.8251617694728592
ROC	0.6806147897296391	0.6658060845455687	0.6819264277337735
False Positive Rate	0.008102256059230284	0.00018767007600638077	0.008054753701461961

5. Key Performance Indicator Analysis

To further explain the concept and give Humana with actionable insights. It is critical to consider the significance of the trait. As a result, we extracted and plotted the top 100 important features (as determined by the LightGBM model).

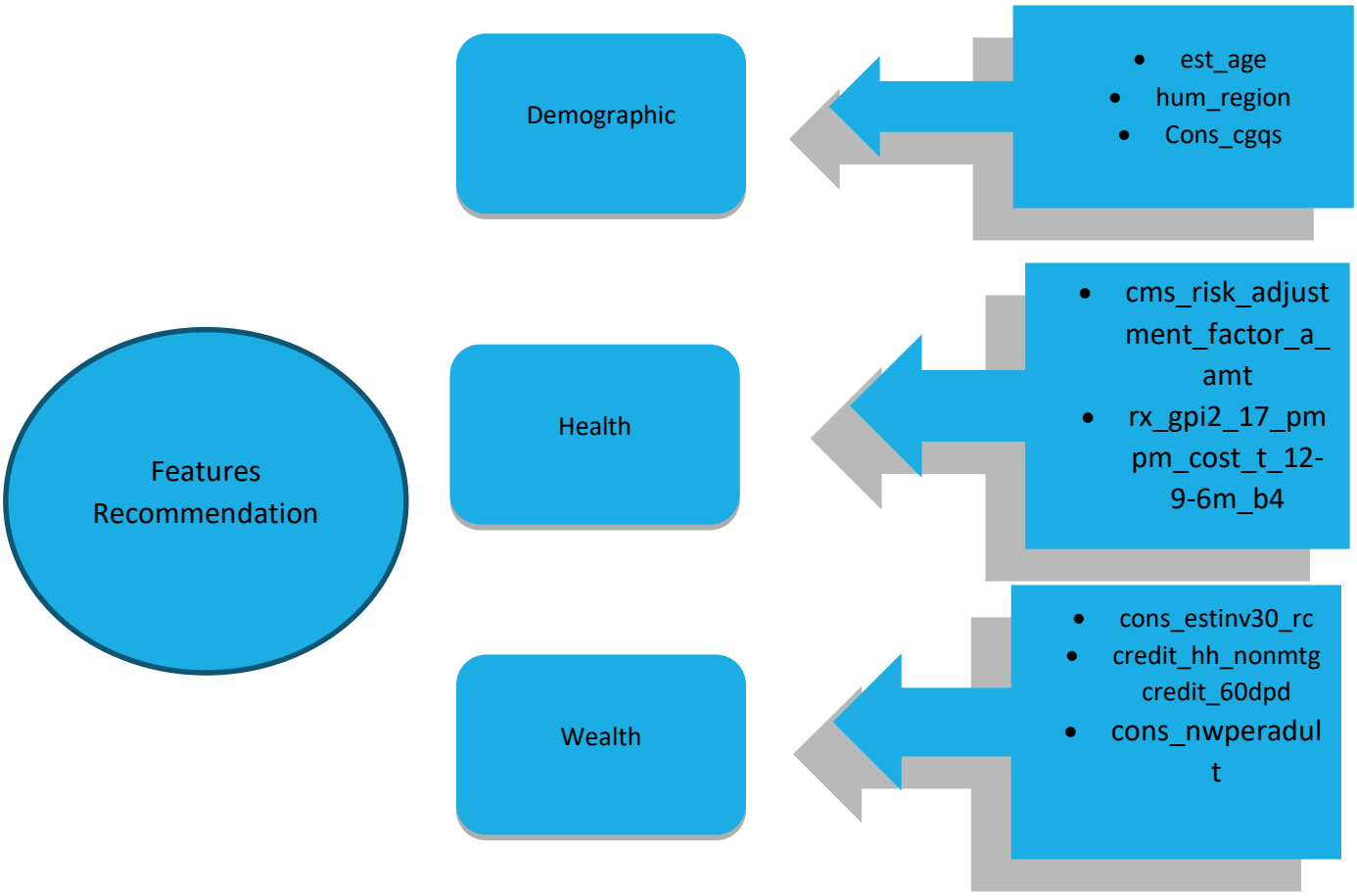


The features listed above fall into several categories, including demographics, CMS, medical claims, credit data, and so on. Because these categories were heavily represented in our top ten features. We will further investigate these qualities and provide recommendations based on our findings. We've included definitions of the ten characteristics for your convenience, which we'll utilize to make additional recommendations.

Feature	**Meaning**
est_age	Member's age
hum_region	Geographic information about the members.
cms_risk_adjustment_factor_a_amt	Amount of risk adjustment factor A
cons_nwperadult	Adults' Net Worth
rx_gpi2_17_pmpm_cost_t_12-9-6m_b4	The trend in the cost per month of prescription associated to VACCINES medications in the sixth to ninth month previous to the score date vs the ninth to twelfth month prior to the score date.
cons_estinv30_rc	Recoded Estimated Household Investable Assets
credit_hh_nonmtgcredit_60dpd	Percentage of non-mortgage loan accounts that are 60 or more days past due.
cons_cgqs	Quality Score for Census Geo-units.

6. Recommendations and Managerial Implications

The models produced a list of factors that contributed significantly to the hesitance score. The sections that follow will analyze these characteristics and suggest ways to increase COVID immunization as needed. Eight of the top fifty most important traits were chosen and divided into three categories: demography, health, and financial considerations.



7. Conclusion

The purpose of this study was to predict who would be resistant to the covid vaccination. We whittled down over 367 characteristics to roughly 320 for usage in the models. The Accuracy of the Catboost model is 0.8251617694728592 which is the best among all the models trained with AUC of 0.684. We were able to separate reluctant members into numerous sub-segments depending on age, geography, health, and money after doing the feature significance analysis. Then we presented actionable information and feasible strategies to assist Humana in promoting vaccinations among these populations. The situation with covid vaccines in the United States has been developing at an unparalleled rate in human history. Humana may, in the future, evolve with the times, adjusting their marketing and PR accordingly, and iterating on this process to perfect their technique.

8. References:

- Adeline, S., Jin, C. H., Hurt, A., Wilburn, T., Wood, D., & Talbot, R. (2021, October 4). Tracking
- https://mays.tamu.edu/humana-tamu-analytics/humana/
- coronavirus around the U.S.: See how your state is doing. NPR. Retrieved October 10, 2021, from
- https://www.npr.org/sections/health-shots/2020/09/01/816707182/map-tracking-the-spread-of-thecoronavirus-in-the-u-s.
- Advanced topics. Advanced Topics - LightGBM 3.3.0.99 documentation. (n.d.). Retrieved October 10,
- 2021, from https://lightgbm.readthedocs.io/en/latest/Advanced-Topics.html#categorical-feature-support.
- Elflein, J. (2021, October 6). U.S. covid-19 cases by day. Statista. Retrieved October 7, 2021, from
- https://www.statista.com/statistics/1103185/cumulative-coronavirus-covid19-cases-number-us-by-day/.
- Least educated states 2021. (n.d.). Retrieved October 10, 2021, from
- https://worldpopulationreview.com/state-rankings/least-educated-states.
- Mayo Foundation for Medical Education and Research. (n.d.). U.S. COVID-19 vaccine tracker: See your
- state's progress. Mayo Clinic. Retrieved October 10, 2021, from
- https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker.
- Ritchie, H., & Mathieu, E. (2020, March 5). Coronavirus (COVID-19) vaccinations - statistics and
- research. Our World in Data. Retrieved October 10, 2021, from https://ourworldindata.org/covidvaccinations?country=USA.
- Smith, M., Heyward, G., & Kasakove, S. (2021, June 28). Why young adults are among the biggest
- barriers to mass immunity. The New York Times. Retrieved October 8, 2021, from
- https://www.nytimes.com/2021/06/28/us/covid-vaccine-immunity.html.
- World Health Organization. (n.d.). Coronavirus disease (COVID-19). World Health Organization.
- Retrieved October 7, 2021, from https://www.who.int/health-topics/coronavirus#tab=tab_1.