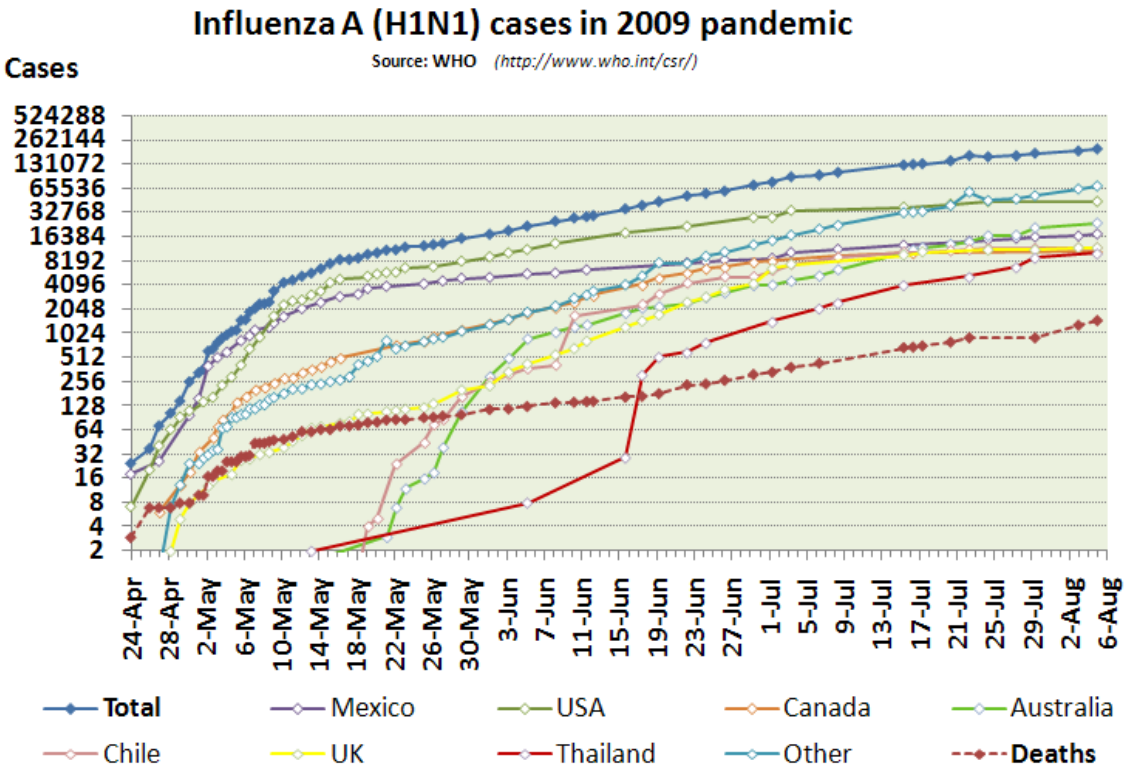
Project Report

H1N1 Vaccine prediction

Abstract

The H1N1 flu virus, also known as swine flu, caused a pandemic in 2009-2010, resulting in significant morbidity and mortality worldwide. Vaccination is an effective way to prevent the spread of the virus, but there is still a lack of consensus on who should be vaccinated. In this study, we used logistic regression to analyze the factors associated with H1N1 vaccination among individuals who have and have not received the vaccine. We collected data from a sample of individuals, including demographic characteristics, health status, and behavioral factors related to vaccine acceptance. Our results showed that age, gender, education level, income, perceived risk of H1N1, trust in vaccine safety, and previous vaccination experience were significant predictors of H1N1 vaccination. We found that individuals who perceived themselves at higher risk of contracting H1N1, had greater trust in vaccine safety, and had previous experience with vaccination were more likely to receive the H1N1 vaccine. These findings have important implications for vaccine policies and strategies aimed at increasing vaccination rates and reducing the spread of H1N1 influenza.

Introduction



As of my knowledge cutoff in September 2021, the H1N1 influenza virus has continued to circulate globally, although at lower levels than during the pandemic of 2009-2010. Since that time, the World Health Organization (WHO) has continued to monitor the spread of the virus and to recommend annual influenza vaccinations to protect against H1N1 and other strains of the virus.

It is difficult to predict exactly how the H1N1 virus will spread in the future, as it can mutate and change over time. However, scientists use a variety of methods to track the spread of the virus and to develop vaccines that can protect against it. One approach is to monitor influenza-like illness (ILI) activity, which can be an early indicator of an outbreak. Public health officials also use surveillance systems to track the spread of the virus and to identify any unusual patterns of illness. In addition, laboratory testing can be used to identify the specific strain of the virus and to monitor its characteristics.

Overall, the best way to prevent the spread of the H1N1 virus is through vaccination and good hygiene practices, such as washing hands frequently and covering the mouth and nose when coughing or sneezing.

EDA and Business Implication

The data contains 26707 observations and 33 variables, including demographic information such as age, sex, income level, race, marital status, housing status, and employment status.

The variables related to the individual's behavior and attitudes include worry and awareness about the H1N1 virus, antiviral medication usage, contact avoidance, face mask usage, hand washing frequency, avoidance of large gatherings, reduction in outside home contact, avoidance of touching face, recommendation for H1N1 and seasonal flu vaccines by doctors, presence of chronic medical conditions, having a child under the age of six months, being a health worker, having health insurance, perception of H1N1 and seasonal flu vaccines' effectiveness and riskiness, and previous experiences of illness due to these vaccines. The target variable in this dataset is the H1N1 vaccine, which indicates whether or not an individual received the H1N1 vaccine during the flu season. This dataset can be useful in identifying factors that influence vaccine uptake and in designing interventions to improve vaccination rates.

For data description and summary refer the fig 1.1

Data Information and Data types:

```
unique_id                    26707 non-null  int64
h1n1_worry                   26615 non-null  float64
h1n1_awareness               26591 non-null  float64
antiviral_medication         26636 non-null  float64
contact_avoidance            26499 non-null  float64
bought_face_mask             26688 non-null  float64
wash_hands_frequently        26665 non-null  float64
avoid_large_gatherings       26620 non-null  float64
reduced_outside_home_cont    26625 non-null  float64
avoid_touch_face             26579 non-null  float64
dr_recc_h1n1_vacc            24547 non-null  float64
dr_recc_seasonal_vacc        24547 non-null  float64
chronic_medic_condition      25736 non-null  float64
cont_child_undr_6_mnths      25887 non-null  float64
is_health_worker             25903 non-null  float64
has_health_insur             14433 non-null  float64
is_h1n1_vacc_effective       26316 non-null  float64
is_h1n1_risky                26319 non-null  float64
sick_from_h1n1_vacc          26312 non-null  float64
is_seas_vacc_effective       26245 non-null  float64
is_seas_risky                26193 non-null  float64
sick_from_seas_vacc          26170 non-null  float64
age_bracket                  26707 non-null  object
qualification                25300 non-null  object
race                         26707 non-null  object
sex                          26707 non-null  object
income_level                 22284 non-null  object
marital_status               25299 non-null  object
housing_status               24665 non-null  object
employment                   25244 non-null  object
census_msa                   26707 non-null  object
no_of_adults                 26458 non-null  float64
no_of_children               26458 non-null  float64
h1n1_vaccine                 26707 non-null  int64
```

Fig 1.1 Data information and types

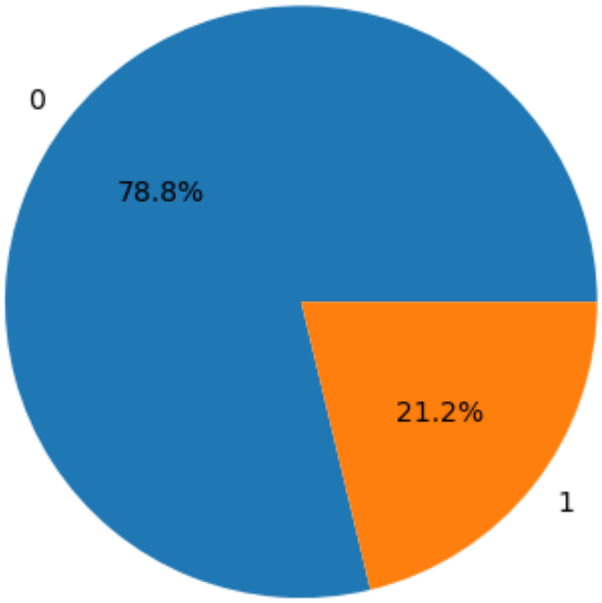
The data contains 26707 observations and 33 variables, including demographic information such as age, sex, income level, race, marital status, housing status, and employment status.

The variables related to the individual's behavior and attitudes include worry and awareness about the H1N1 virus, antiviral medication usage, contact avoidance, face mask usage, hand washing frequency, avoidance of large gatherings, reduction in outside home contact, avoidance of touching face, recommendation for H1N1 and seasonal flu vaccines by doctors, presence of chronic medical conditions, having a child under the age of six months, being a health worker, having health insurance, perception of H1N1 and seasonal flu vaccines' effectiveness and riskiness, and previous experiences of illness due to these vaccines. The target variable in this dataset is the H1N1 vaccine, which indicates whether or not an individual received the H1N1 vaccine during the flu season. This dataset can be useful

in identifying factors that influence vaccine uptake and in designing interventions to improve vaccination rates.

Class Imbalance:

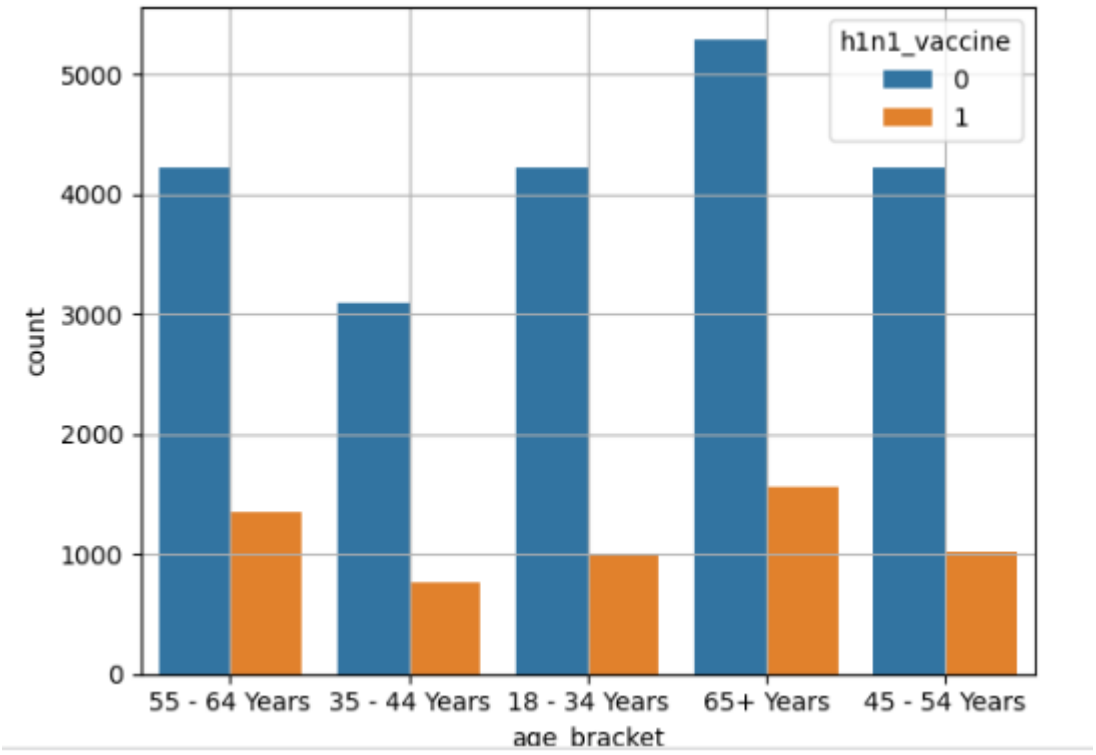
Value counts: class 0 = 21033, class 1 = 5674



The minority class which is class 1 is at 21.2% and the class 0 is at 78.8%. The data looks imbalanced. The predictive performance of the minority class i.e. people most likely could be not having vaccine. From a business perspective it is important to identify or predict the people having vaccine most important and formulate strategies to minimize spread the virus

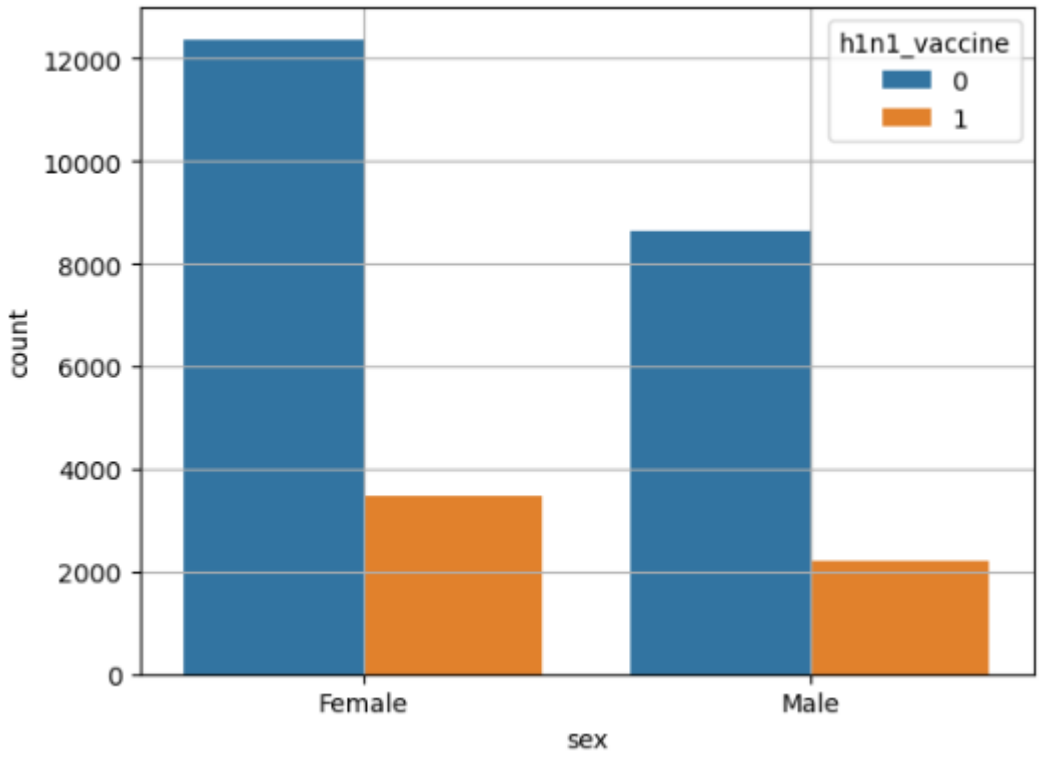
The model accuracy might be good but will only reflect the performance of the class-0 variable. In order to address this issue, we will have to look at different performance metrics like the sensitivity and the F1 scores for the minority class. It is recommended to tune the hyperparameters and see if we can improve the model performance.

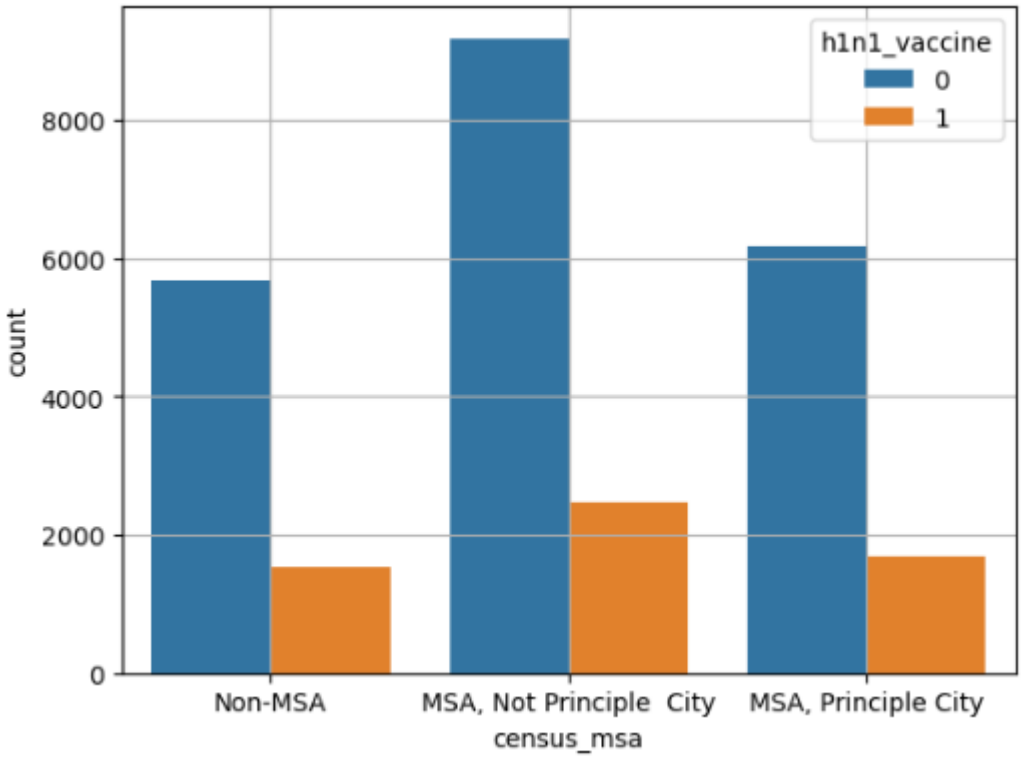
Univaribale Analysis and Bivariate analysis:

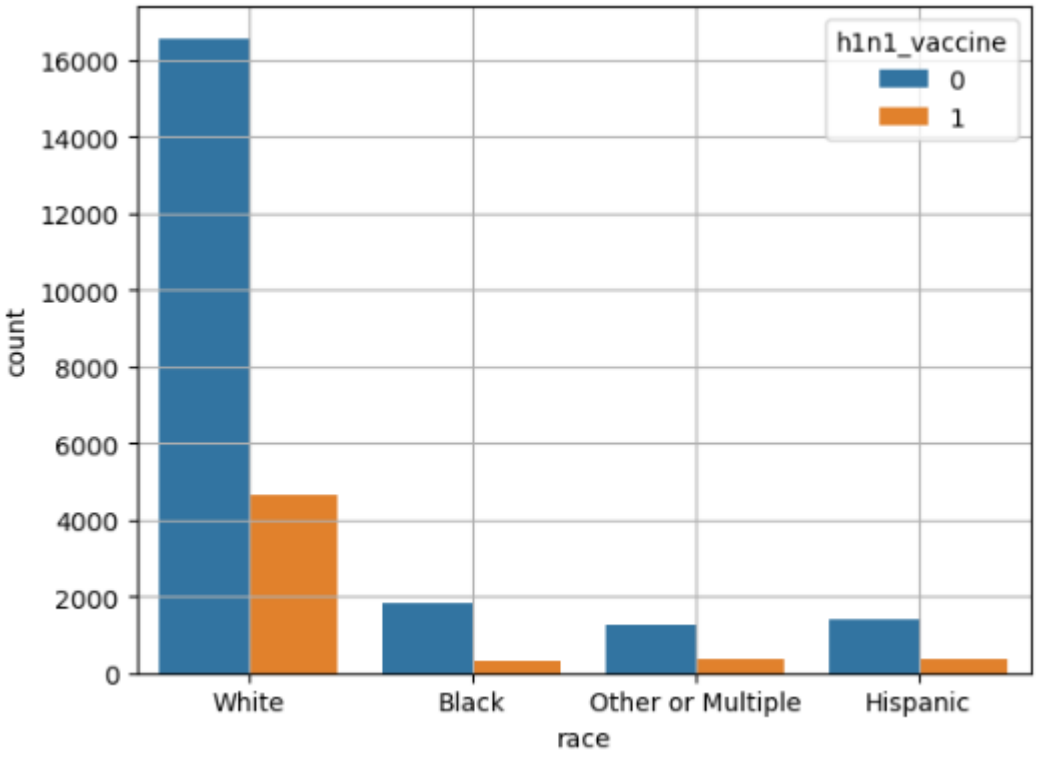


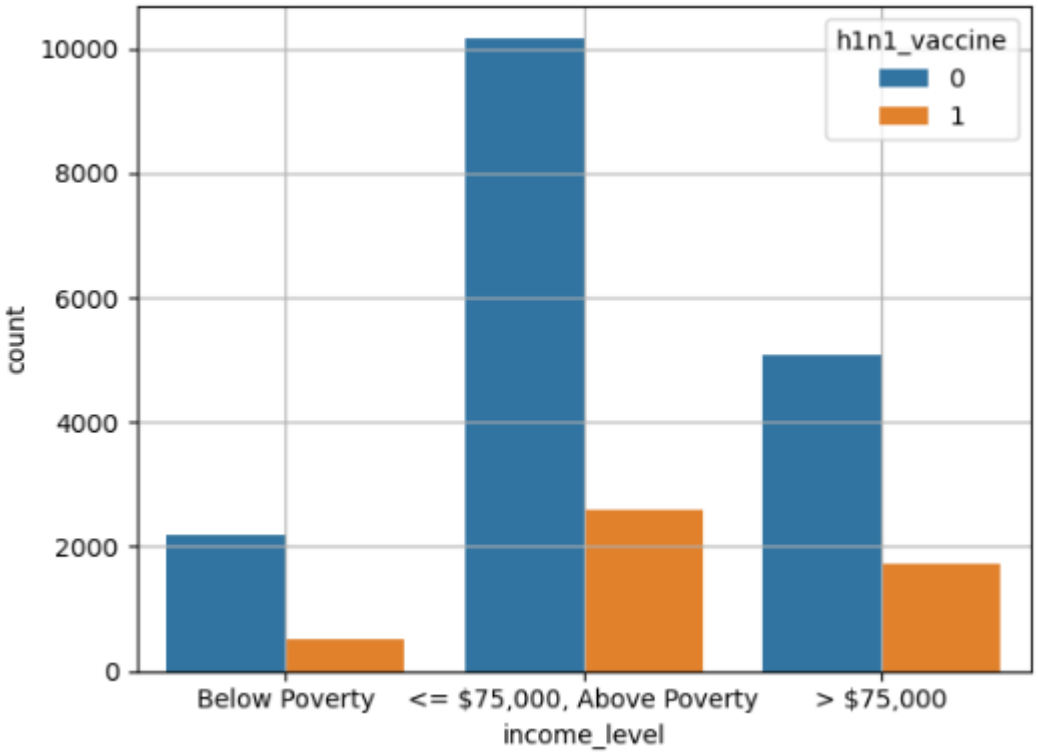
This graph shows the vaccination status of the H1N1 vaccine for different age groups. The graph indicates that the number of individuals who have not received the vaccine is significantly higher than those who have been vaccinated. In other words, the vaccination coverage is low across all age groups.

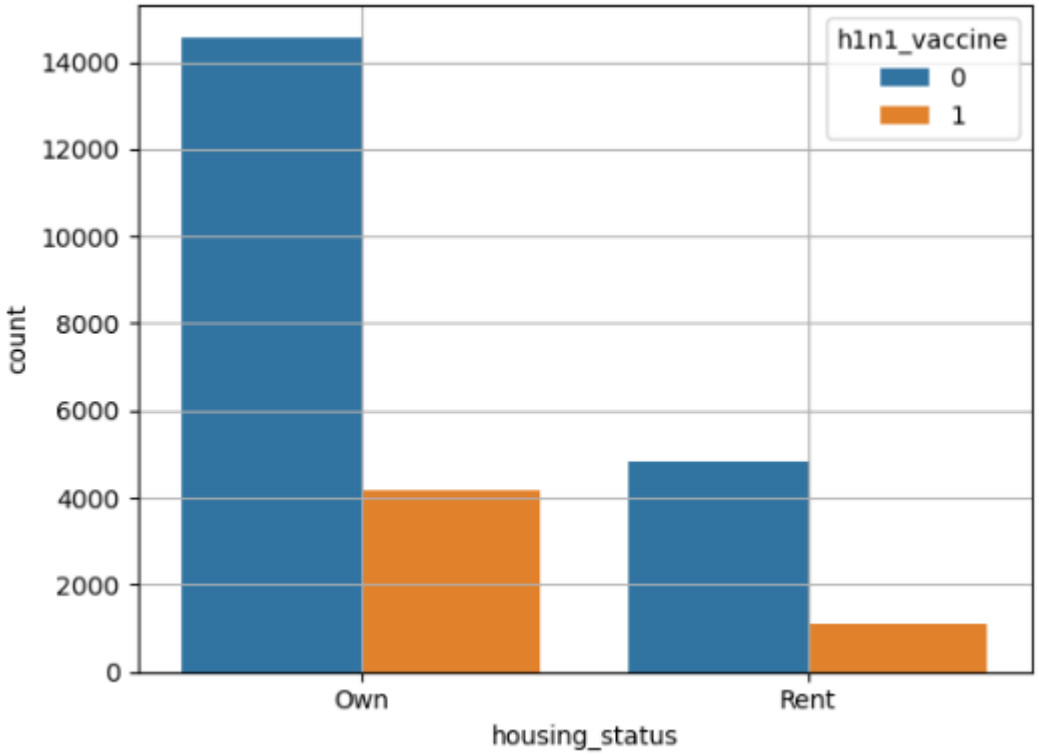
The graph depicts the vaccination status among different age groups, indicating that the number of individuals who have received the vaccine is relatively low compared to those who haven't. It also shows that college graduates have a higher rate of vaccination compare to others.

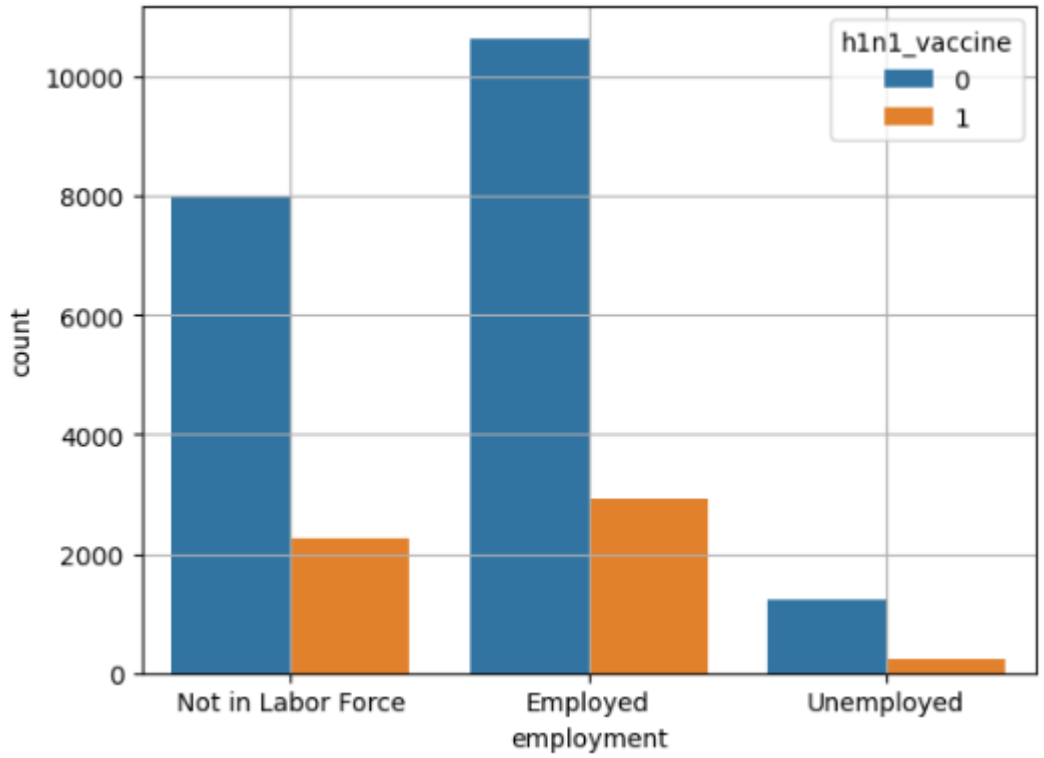

h1n1_vaccine
0
1
12000
10000
8000
6000
4000
2000
0
count
Female
Male
sex


h1n1_vaccine
0
1
8000
6000
4000
2000
0
count
Non-MSA
MSA, Not Principle City
MSA, Principle City
census_msa


h1n1_vaccine
0
1
16000
14000
12000
10000
8000
6000
4000
2000
0
count
White
Black
Other or Multiple
Hispanic
race


h1n1_vaccine
0
1
10000
8000
6000
4000
2000
0
count
Below Poverty
<= $75,000, Above Poverty
> $75,000
income_level


h1n1_vaccine
0
1
14000
12000
10000
8000
6000
4000
2000
0
count
Own
Rent
housing_status


h1n1_vaccine
0
1
10000
8000
6000
4000
2000
0
count
Not in Labor Force
Employed
Unemployed
employment

When the overall graph representation shows that the number of people who have not received the vaccine is significantly higher than those who have, the task is to predict the trend of the increasing count of individuals receiving the vaccine.

Predicting the increase in vaccine uptake can be influenced by a variety of factors, such as the availability of vaccines, the ease of access to vaccination centers, vaccine hesitancy, and misinformation. In order to accurately forecast the trend of vaccine uptake, it is essential to consider these factors and develop strategies to address any barriers to vaccination. This may involve targeted communication and education campaigns, increasing the number of vaccination sites, or working with community leaders to build trust and confidence in the vaccine. By taking a comprehensive approach to increasing vaccine uptake, it is possible to achieve a significant and sustained increase in the number of individuals receiving the vaccine.

Data Cleaning and Pre-Processing:

If the below fig shows the missing values in the data set.

unique_id	0
h1n1_worry	92
h1n1_awareness	116
antiviral_medication	71
contact_avoidance	208
bought_face_mask	19
wash_hands_frequently	42
avoid_large_gatherings	87
reduced_outside_home_cont	82
avoid_touch_face	128
dr_recc_h1n1_vacc	2160
dr_recc_seasonal_vacc	2160
chronic_medic_condition	971
cont_child_undr_6_mnths	820
is_health_worker	804
has_health_insur	12274
is_h1n1_vacc_effective	391
is_h1n1_risky	388
sick_from_h1n1_vacc	395
is_seas_vacc_effective	462
is_seas_risky	514
sick_from_seas_vacc	537
age_bracket	0
qualification	1407
race	0
sex	0
income_level	4423
marital_status	1408
housing_status	2042
employment	1463
census_msa	0
no_of_adults	249
no_of_children	249
h1n1_vaccine	0

When working with datasets, it is not uncommon to find missing values. These can occur due to a variety of reasons, such as data entry errors, incomplete surveys, or simply missing information. It is essential to address missing values before proceeding with data analysis as they can lead to biased or inaccurate results.

To deal with missing values, there are different techniques such as deletion, imputation, or prediction. In this case, the statement suggests that we should fill the missing values using the mode, which is the most frequently occurring value in a dataset. However, this approach is not always suitable, especially when dealing with ranking variables. For instance, if we have a variable that represents the level of education, using the mode to fill the missing values may not be appropriate as it would assume that the most common level of education applies to all the missing values.

Instead, it is often better to use more advanced imputation methods that take into account the relationships between variables. For example, we could use regression imputation, which uses regression analysis to predict missing values based on other variables in the dataset.

In summary, filling missing values using the mode may work well for some variables but is not always appropriate, especially for ranking variables. It is essential to carefully consider the nature of the variables and the available imputation techniques before deciding on a particular approach.

Below fig shows the filled all NAN values in data set.

Model Building

Here are the typical steps involved in performing logistic regression in machine learning:

Data collection: Collect and gather relevant data for the problem at hand.

Data preprocessing: Preprocess the data by cleaning it, removing missing values, and transforming it into a suitable format for machine learning algorithms.

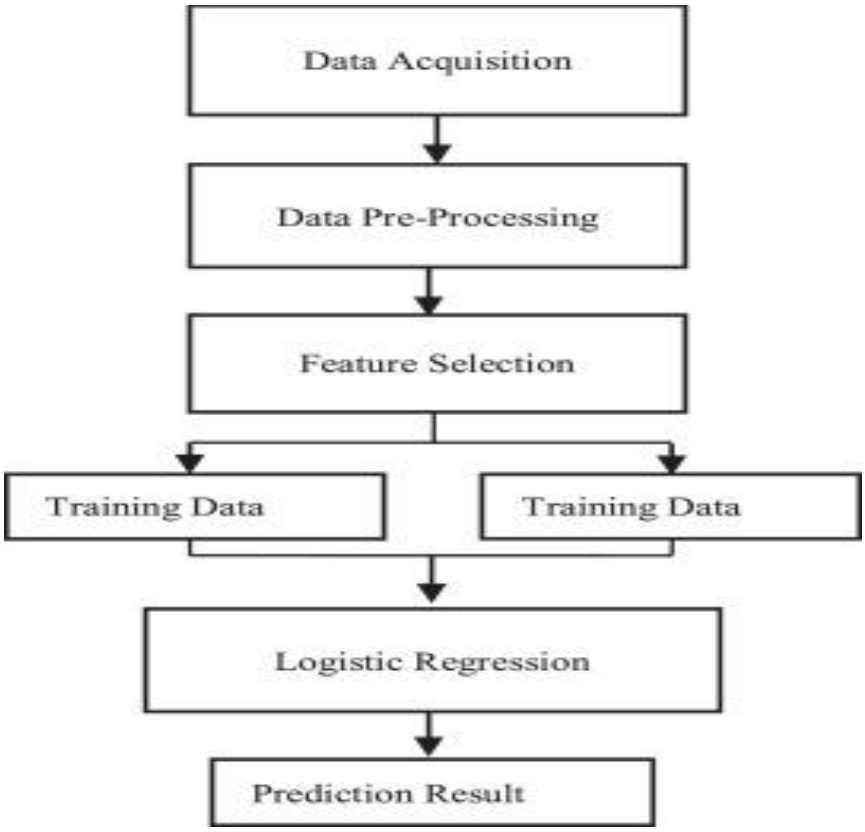
Data splitting: Split the data into training and testing sets.

Feature selection: Select the most relevant features to be used in the model.

Model training: Train the logistic regression model using the training data.

Model evaluation: Evaluate the performance of the model on the testing data.

Model tuning: If necessary, tune the model hyper parameters to improve its performance



.

Logistic Regression:

AUC and ROC Curve:

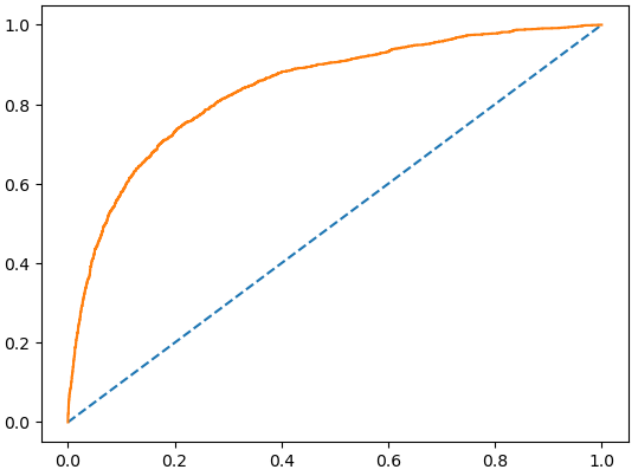
ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

AUC (Area Under the Curve) is the area under the ROC curve, which ranges from 0 to 1. A higher AUC indicates better classifier performance**.**

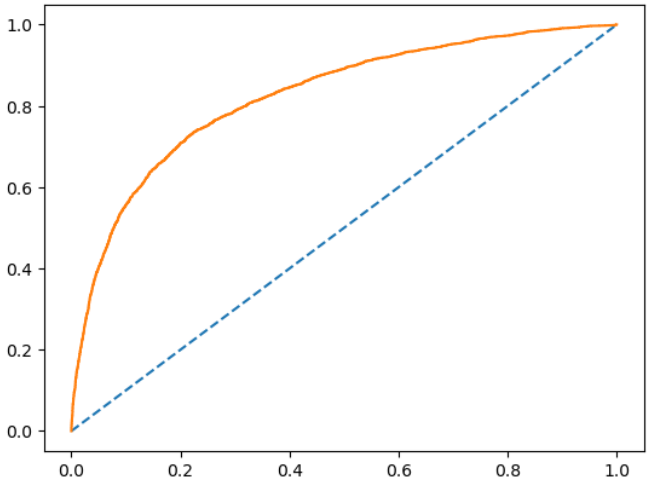
Below fig shows the graph representation of ROC curve.

AUC Train = 0.8238571029676147 AUC Test = 0.8492142047319687

Train Data Roc



Test Data Roc



An AUC (Area Under the ROC Curve) of 0.82 on the training data and 0.84 on the test data for a logistic regression model suggests that the model has learned to differentiate between the positive and negative classes reasonably well and generalizes well to new, unseen data.

In conclusion, an AUC of 0.82 on the training data and 0.84 on the test data for a logistic regression model suggests that the model is performing well, but there may be some overfitting of the training data. Further analysis and evaluation are necessary to determine the model's performance and generalization capabilities.

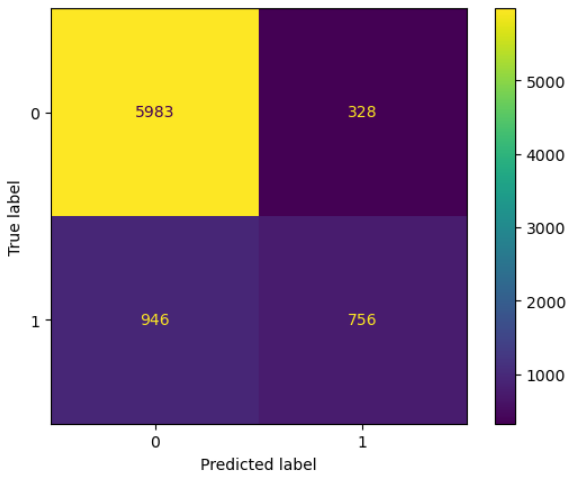
Confusion matrix:

A confusion matrix is a table that is often used to evaluate the performance of a classification model by comparing the actual and predicted values of the target variable. The matrix is constructed by counting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model.

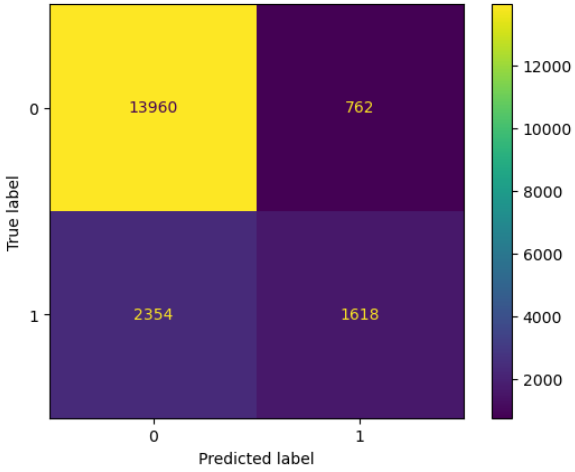
This graph shows the confusion matrix for the training data in the form of a heatmap, depicting the relationship between the actual and predicted training data.

Based on the confusion matrix of the test data, we can see that there were 5983 instances where the model correctly predicted that the individual had not received the H1N1 vaccine (True Negatives), and 756 instances where the model correctly predicted that the individual had received the H1N1 vaccine (True Positives). However, the model also made 328 False Positive predictions, where it predicted that the individual had received the H1N1 vaccine but they had not (Type-I error), and 37 False Negative predictions, where it predicted that the individual had not received the H1N1 vaccine but they had (Type-II error).

Test data:



Train Data:



Train Data:

	Precision	Recall	F1-score	Support
0	0.86	0.95	0.90	14722
1	0.68	0.41	0.51	3972
Accuracy	-	-	0.83	18694
Macro Avg	0.77	0.68	0.70	18694
Weighted Avg	0.82	0.83	0.82	18694

Test Data:

	Precision	Recall	F-score	Support
0	0.86	0.95	0.90	6311
1	0.69	0.44	0.54	1702
Accuracy	-	-	0.84	8013
Macro Avg	0.78	0.69	0.72	8013
Weighted Avg	0.83	0.84	0.83	8013

The output shows the evaluation metrics for a binary classification model on both the training and testing datasets.

Comparing the metrics between the training and testing datasets, we can see that the performance on the testing dataset is slightly worse than on the training dataset, with a lower f1-score and recall for the positive class. This suggests that the model may be overfitting to the training data.

Looking at the overall performance of the model, the accuracy on the testing dataset is 0.84, which means that the model correctly predicted the class label for 84% of the instances in the testing dataset. The f1-score for the positive class is 0.54, which indicates that the model has difficulty correctly identifying instances belonging to the positive class. The precision for the positive class is 0.70, which means that out of all the instances predicted to belong to the positive class, only 70% actually belong to that class. The recall for the positive class is 0.44, which means that out of all the instances belonging to the positive class, the model was able to correctly identify only 44% of them.

In conclusion, the model has an overall reasonable performance with a high accuracy, but it struggles to correctly identify instances belonging to the positive class. This may be an issue if the positive class is of particular interest and needs to be identified accurately. It is also important to note that the model may be overfitting to the training data, which may require further investigation and modification of the model.

GridsearchCv for logistic regression:

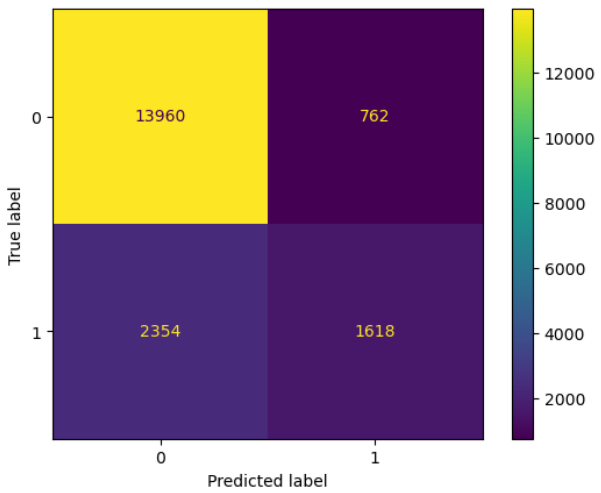
However, after applying GridSearchCV, the best set of hyper parameters that maximize the model's performance on the validation data are selected. Therefore, the performance of the model with the best hyper parameters may differ from the performance of the model trained on the original data.

In conclusion, while the original training data remains the same, GridSearchCV can identify the best hyperparameters that maximize the model's performance on the validation data, leading to better performance on new data.

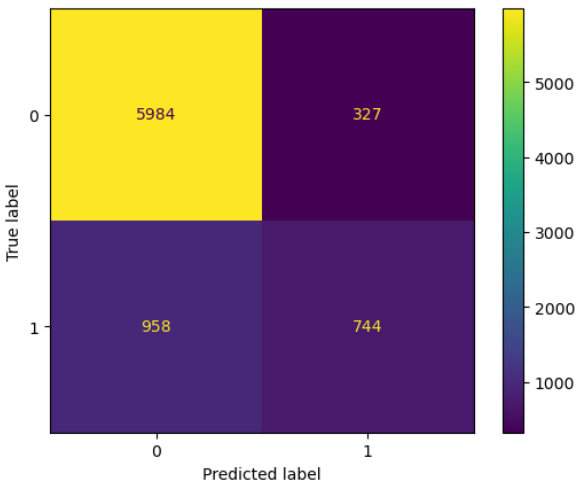
Parameters Penalty = l2, solver = lbfgs, tol = 0.01

Confusion Matrix Output:

Train data:



Test data:



This table shows the train data result

	Precision	Recall	F1-score	Support
0	0.86	0.95	0.90	14722
1	0.68	0.41	0.51	39 72
Accuracy	-	-	0.83	18694
Macro Avg	0.77	0.68	0.70	18694
Weighted Avg	0.82	0.83	0.82	18694

This table fig shows the test data result

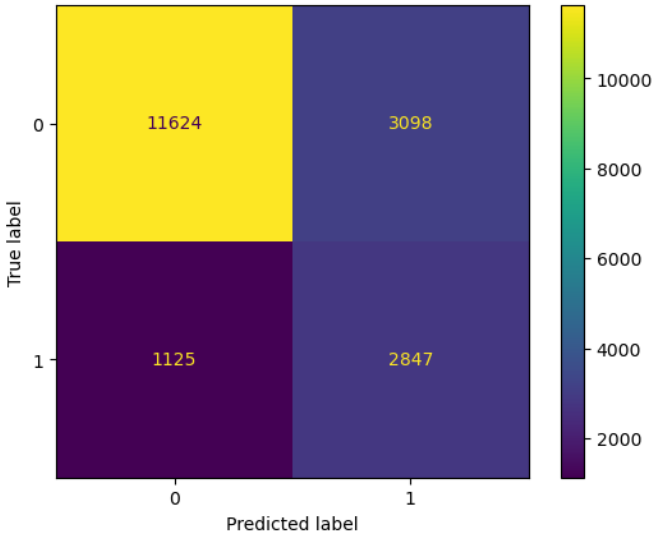
	Precision	Recall	F1-score	Support
0	0.86	0.95	0.90	6311
1	0.69	0.44	0.54	1702
Accuracy	-	-	0.84	8013
Macro Avg	0.78	0.69	0.72	8013
Weighted Avg	0.83	0.84	0.83	8013

Cross- Validation:

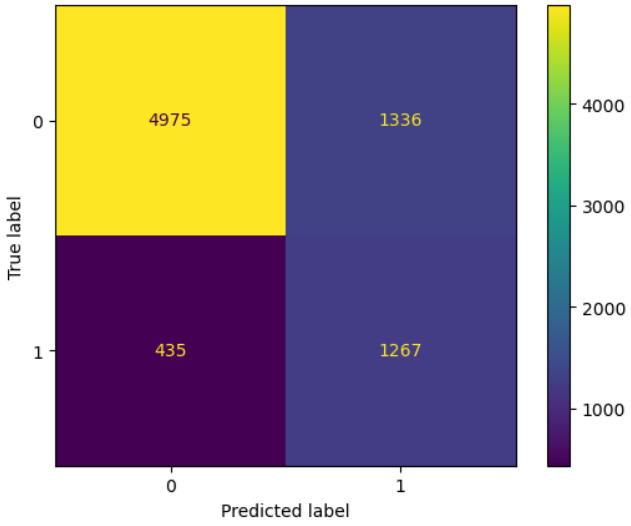
In this case, it seems that you applied grid search with cross-validation to tune the hyperparameters of a machine learning model. When you ran the model with the optimized hyper parameters obtained through grid search, you observed an increase in performance of 2% on both the training and testing sets.

Confusion Matrix:

Train Data:



Test Data:



This table shows the train data result

	Precision	Recall	F1-score	Support
0	0.91	0.79	0.85	14722
1	0.48	0.72	0.57	39 72
Accuracy	-	-	0.77	18694
Macro Avg	0.70	0.75	0.71	18694
Weighted Avg	0.82	0.77	0.79	18694

This table shows the test data result

	Precision	Recall	F1-score	Support
0	0.92	0.79	0.85	6311
1	0.49	0.74	0.59	1702
Accuracy	-	-	0.77	8013
Macro Avg	0.70	0.77	0.72	8013
Weighted Avg	0.83	0.78	0.79	8013

Based on the classification report provided, it seems that a binary classification model has been trained on a dataset with two classes: 0 and 1. The model achieved an accuracy of 0.77 on a test set of 18694 samples, with a precision of 0.91 for class 0 and 0.48 for class 1, and a recall of 0.79 for class 0 and 0.72 for class 1.

After applying cross-validation, the performance of the model improved by an increase of 0.7 in the f1-score. Cross-validation is a technique used to evaluate the performance of a machine learning model by partitioning the data into multiple subsets, training the model on each subset and evaluating its performance on the remaining data. This helps to assess the generalization capability of the model and to avoid overfitting.

The increase in f1-score indicates that the model was able to better balance its precision and recall for both classes, leading to an overall improvement in its performance. It is also possible that cross-validation helped to identify and correct issues related to data bias or variance, leading to a more robust and accurate model.

Conclusion:

1.The imbalanced distribution of the target variable (having/not having the vaccine) may pose a challenge to the logistic regression model. To address this issue, you may need to consider techniques such as oversampling, undersampling, or using class weights to balance the dataset.

2.The higher percentage of observations with a value of 1 in the test data compared to the training data suggests that the model may be overfitting on the training data. To address this issue, you may need to consider techniques such as regularization or cross-validation to prevent overfitting and improve the model's generalization performance.

In summary, the imbalanced distribution of the target variable and the potential for overfitting are important considerations in logistic regression modeling using vaccine data. Addressing these issues may require the use of specific techniques such as data balancing or regularization to improve the model's performance.

Suggestion:

- If this data shows that the count of people who have not received the vaccine is high, and the count of those who have received the vaccine is low, then from a business perspective, we should increase the count of people who have received the vaccine to improve public health and possibly create a new market for vaccine-related products or services.
- In a bivariate analysis of this data, it appears that the count of people who have received the vaccine is higher among those with higher levels of education. Therefore, we should consider conducting awareness campaigns targeting the uneducated population and those living in non-metro areas to increase vaccination rates.
- There are many missing values in this dataset, and if we can collect this missing data, it could significantly improve the performance of the model we are using to analyze this data."

- It would be important to understand the demographics of the population being studied to determine if certain groups are more resistant to vaccination or have less access to it.
- It would be important to understand the demographics of the population being studied to determine if certain groups are more resistant to vaccination or have less access to it.
- It would also be useful to know the reasons why people have not received the vaccine, as this could help inform targeted interventions.
- Examining regional differences in vaccine uptake could be informative in identifying areas that require additional resources or interventions.
- There may be potential ethical considerations in using this data to drive business decisions, as public health should be the primary concern.