PROJECT REPORT

Predict H1N1 vaccine with Logistic Regression¶

By Somashree Sahoo

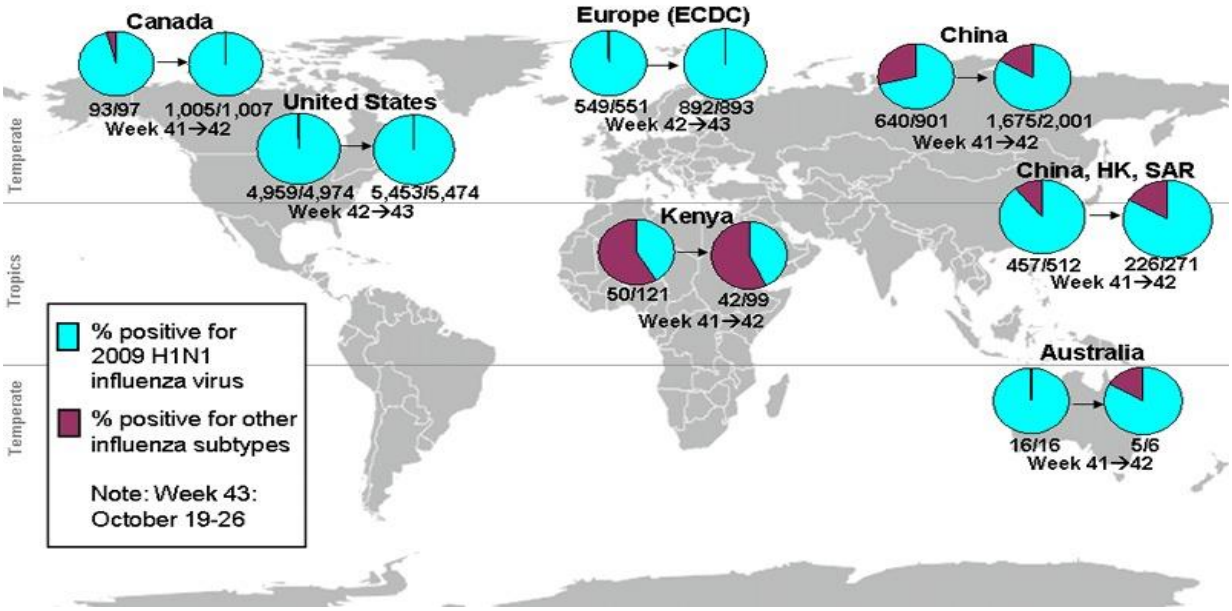
Abstract:

Subjects receiving the same vaccine often show different levels of immune responses and some may even present adverse side effects to the vaccine. Systems vaccinology can combine data and machine learning techniques to obtain highly predictive signatures of vaccine immunogenicity and reactogenicity. Currently, several machine learning methods are already available to researchers with no background in bioinformatics.

Introduction:

THE H1N1 VIRUS

H1N1 or swine flu virus first emerged in the spring of 2009 in Mexico and then in the United States and quickly spread across the globe.



A unique combination of influenza genes was discovered in this novel H1N1 virus which was not identified prior in humans or animals.

This contagious novel virus had a very powerful impact on the whole world on June 11, 2009, the World Health Organization (WHO) declared that a pandemic of 2009 H1N1 flu or swine flu had begun.

According to the CDC, the first and foremost step in protecting oneself of this virus is a yearly flu vaccination. Various factors such as age, the health status of an individual which affects the ability of the vaccination to provide protection to the person who is vaccinated.

Several activities were performed using various social media platforms and broadcasting networks such as Twitter to track the levels of disease activity and the concern of the public towards this pandemic situation.

Problem Statement:

Predict how likely it is that the people will take an H1N1 flu vaccine using Logistic Regression.

Scope:

- Exploratory data analysis
- Data Pre-processing
- Training logistic regression model with MLE for prediction
- Tuning the model to improve the performance

EDA and Business Implication

For data description and summary

RangeIndex: 26707 entries, 0 to 26706
Data columns (total 34 columns):

#	Column	Non-Null Count	Dtype

```
 0   unique id                      26707 non-null int64
 1   h1n1_worry                     26615 non-null float64
 2   h1n1_awareness                 26591 non-null float64
 3   antiviral medication           26636 non-null float64
 4   contact avoidance              26499 non-null float64
 5   bought face mask               26688 non-null float64
 6   wash hands frequently          26665 non-null float64
 7   avoid large gatherings         26620 non-null float64
 8   reduced outside home count     26625 non-null float64
 9   avoid touch face               26579 non-null float64
 10 dr_recc_h1n1_vacc               24547 non-null float64
 11 dr_recc_seasonal_vacc           24547 non-null float64
 12 chronic_medic_condition         25736 non-null float64
 13 cont_child_undr_6_mnths         25887 non-null float64
 14 is_health_worker                25903 non-null  float64
 15 has_health_insur                14433 non-null  float64
 16 is_h1n1_vacc_effective          26316 non-null  float64
 17 is_h1n1_risky                   26319 non-null  float64
 18 sick_from_h1n1_vacc             26312 non-null  float64
 19 is_seas_vacc_effective          26245 non-null  float64
 20 is_seas_risky                   26193 non-null  float64
 21 sick_from_seas_vacc             26170 non-null  float64
 22 age_bracket                     26707 non-null  object
 23 qualification                   25300 non-null object
 24 race                            26707 non-null  object
 25 sex                             26707 non-null  object
 26 income level                    22284 non-null  object
 27 marital_status                  25299 non-null  object
 28 housing status                  24665 non-null object
 29 employment                      25244 non-null object
 30 census_msa                      26707 non-null object
 31 no_of_adults                    26458 non-null float64
 32 no_of_children                  26458 non-null float64
 33 h1n1_vaccine                    26707 non-null int64
dtypes: float64(23), int64(2), object (9)
```

Several features including a persons' behavior' and his/her 'opinion' about the vac cine, marital status, geographic location all are found, so the data consists a signifi cantly large amount of info.

Data information and types:

The data types are appropriately listed. Unique id and h1n1 vaccine are ordinal dat a. Age, qualification, race, sex, income level, marital status, housing status, employ ment, census msa are object data types and the rest are numerical float data types . We could also observe the presence of Null values. There are no unwanted punct uation marks, spaces, prefixes or suffixes. The existing unique id, has health insura nce might not be necessary so we are dropping the columns.

Checking the Null Values:

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
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

dtype: int64

Seeing the categorical features present in the data:

	age_bracket	qualification	race	sex	income_level	marital_status	housing_status	employment	census_msa
0	55 - 64 Years	< 12 Years	White	Female	Below Poverty	Not Married	Own	Not in Labor Force	Non-MSA
1	35 - 44 Years	12 Years	White	Male	Below Poverty	Not Married	Rent	Employed	MSA, Not Principle City
2	18 - 34 Years	College Graduate	White	Male	<= $75,000, Above Poverty	Not Married	Own	Employed	MSA, Not Principle City
3	65+ Years	12 Years	White	Female	Below Poverty	Not Married	Rent	Not in Labor Force	MSA, Principle City
4	45 - 54 Years	Some College	White	Female	<= $75,000, Above Poverty	Married	Own	Employed	MSA, Not Principle City

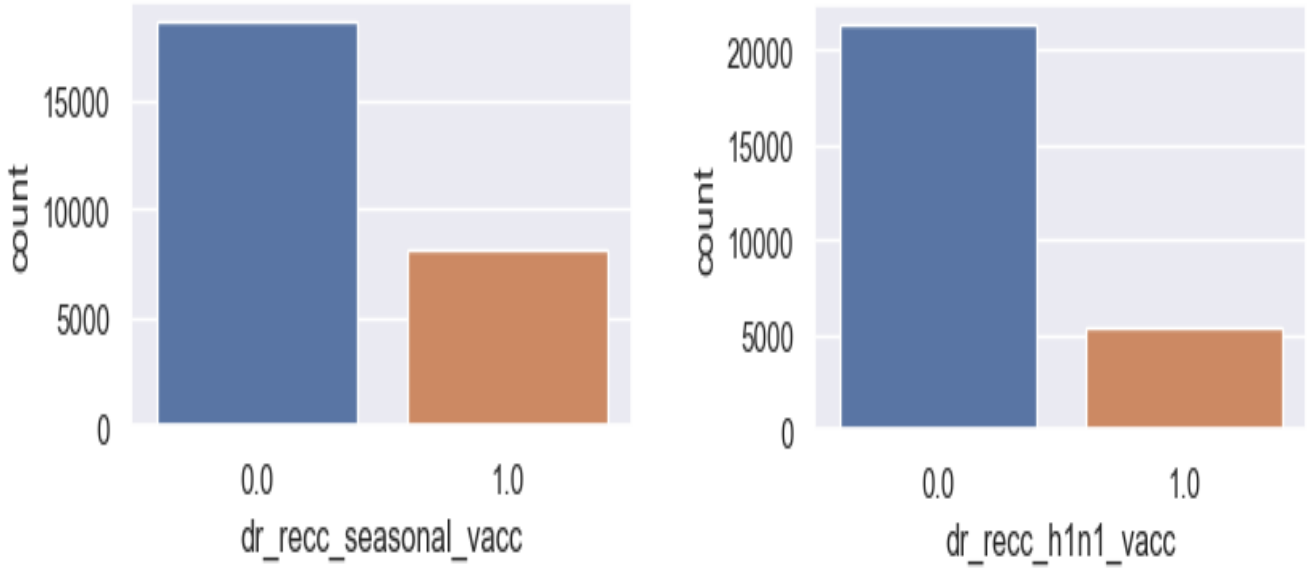
There are 9 categorical features. All Categorical features which are NULL are substituted with the mode of data, and the Numeric features with NULLs are substituted with the median.

Treating the null values of all the categorical and numerical data:

```
h1n1_worry                     0
h1n1_awareness                 0
antiviral_medication           0
contact_avoidance              0
bought_face_mask               0
wash_hands_frequently          0
avoid_large_gatherings         0
reduced_outside_home_cont      0
avoid_touch_face               0
dr_recc_h1n1_vacc              0
dr_recc_seasonal_vacc          0
chronic_medic_condition        0
cont_child_undr_6_mnths        0
is_health_worker               0
is_h1n1_vacc_effective         0
is_h1n1_risky                  0
sick_from_h1n1_vacc            0
is_seas_vacc_effective         0
is_seas_risky                  0
sick_from_seas_vacc            0
age_bracket                    0
qualification                  0
race                           0
sex                            0
income_level                   0
marital_status                 0
housing_status                 0
employment                     0
census_msa                     0
no_of_adults                   0
no_of_children                 0
h1n1_vaccine                   0
dtype: int64
```

Visualizing and analyzing the data:

Analyzing the number of people who took each vaccine:



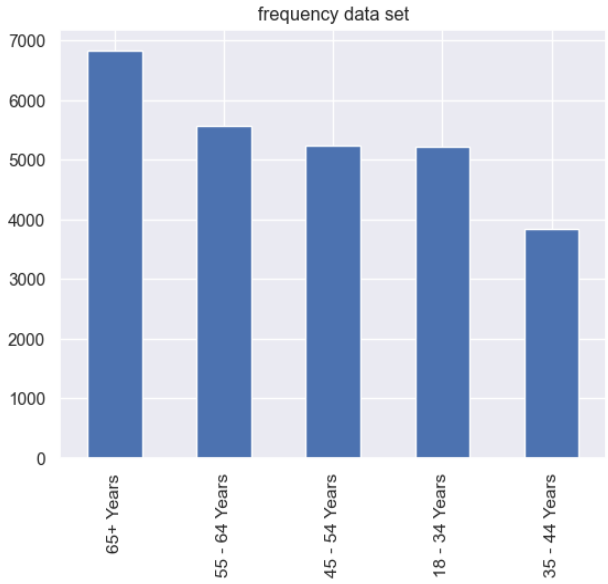
Here 0.0 represents that person didn't take the vaccine and 1.0 reprents the person took the vaccine.

Considerably larger number of people have chosen to take the seasonal flu vaccine compared to those who took the H1N1 vaccine.

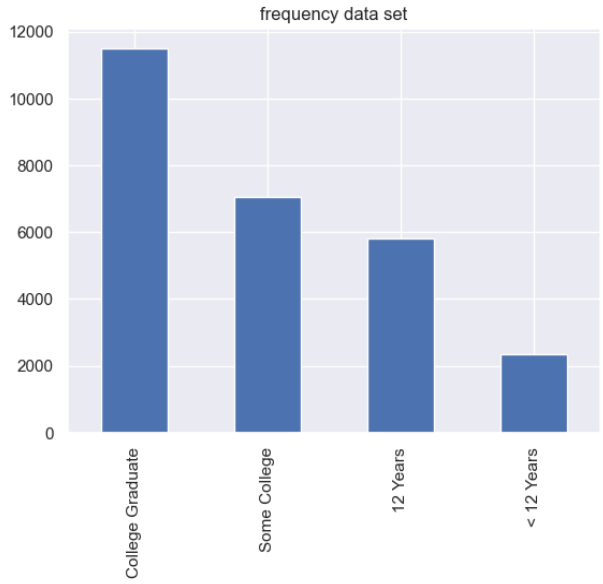
- **There are no outliers as the data's are basically in categorical form.**

Univariate Analysis and Bivariate Analysis:

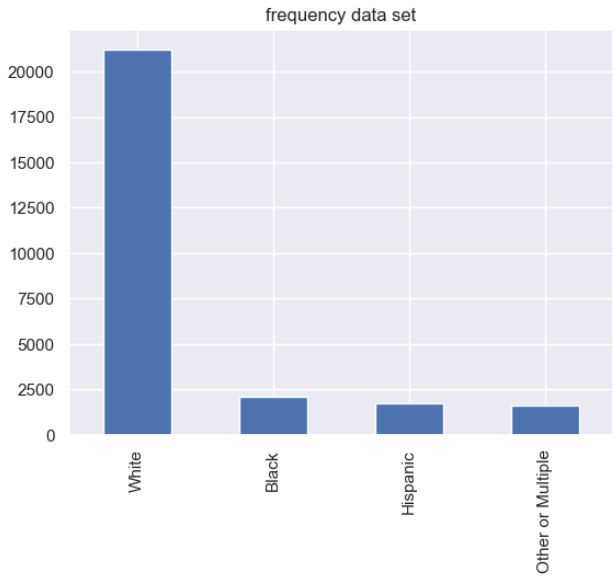
Age bracket



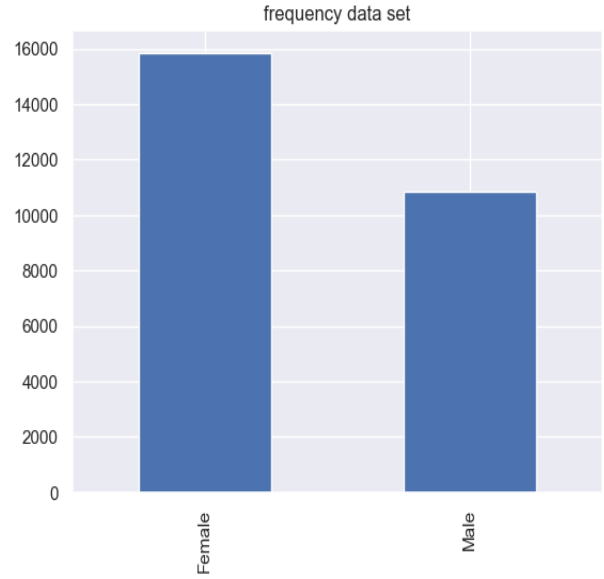
Qualification



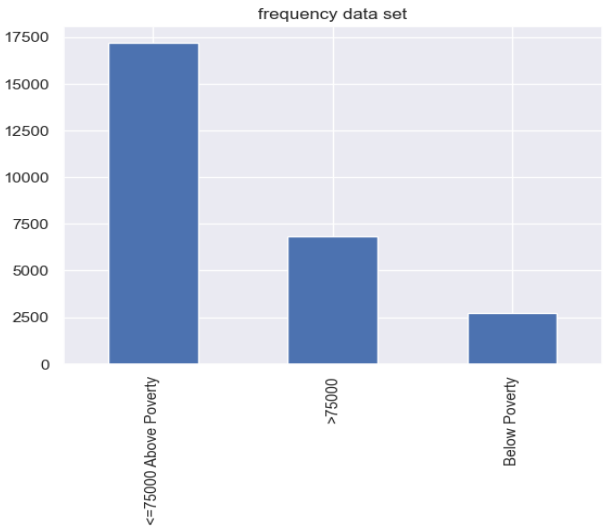
Details of race



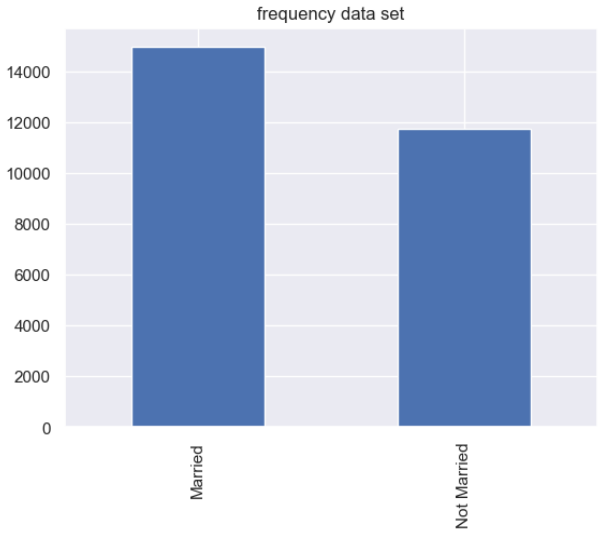
Details of sex



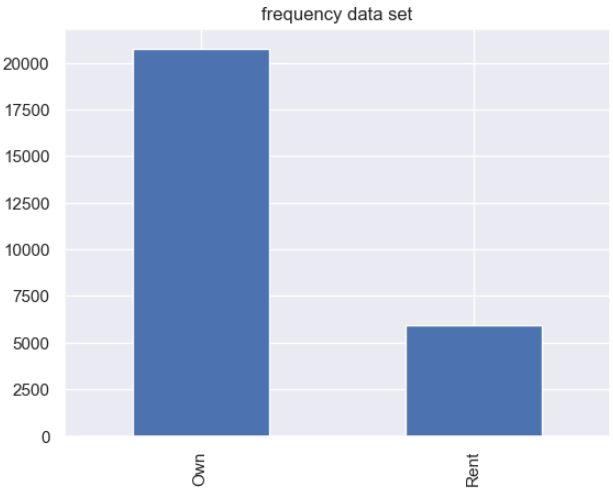
Details of income level



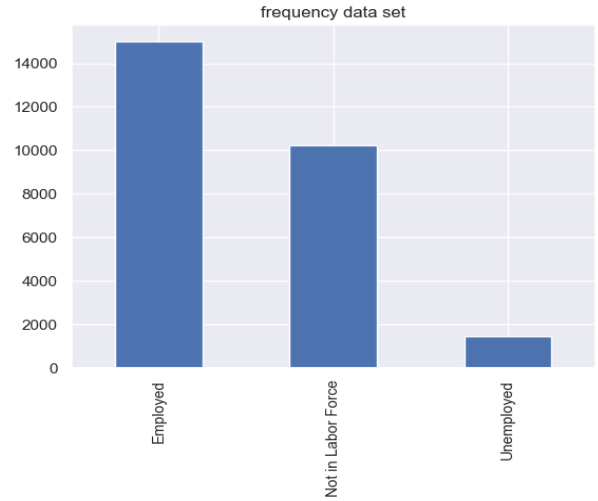
Details of marital status



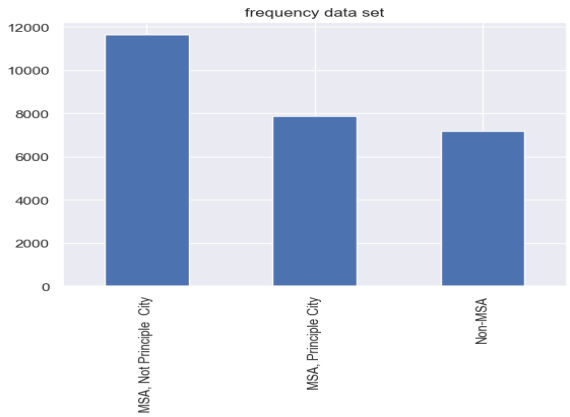
Housing status



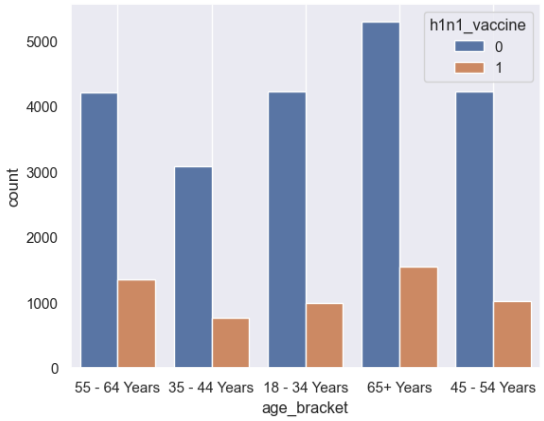
Employment

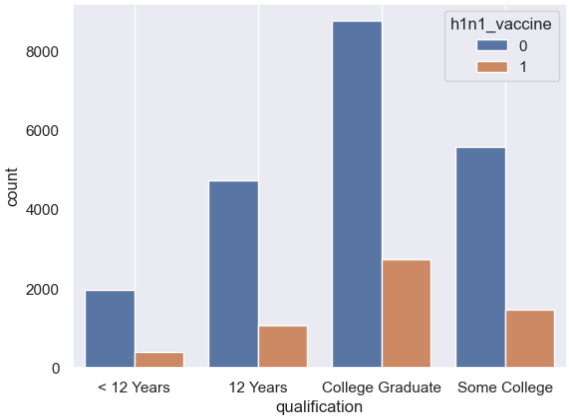


Census msa

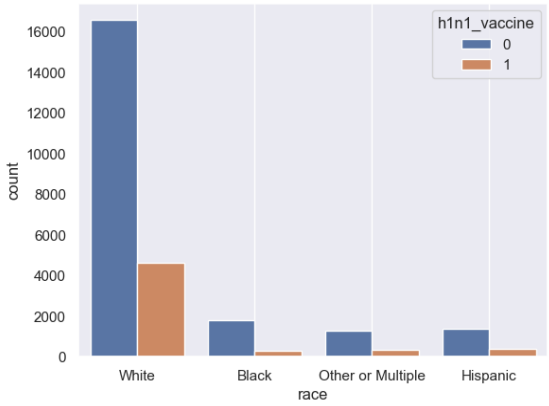


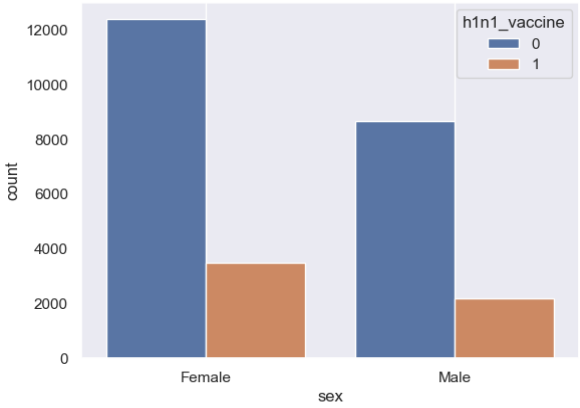
Observing the variety personal information, checking the people who have taken the h1n1 vaccine:



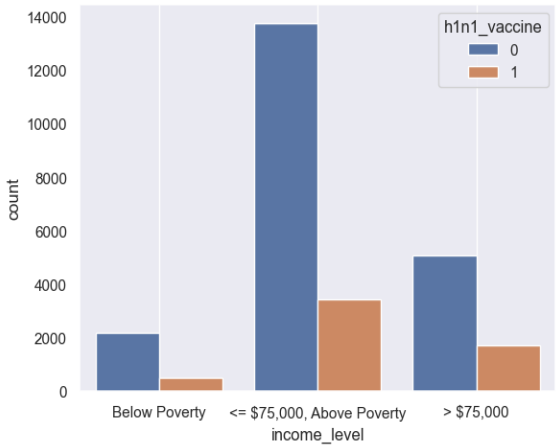


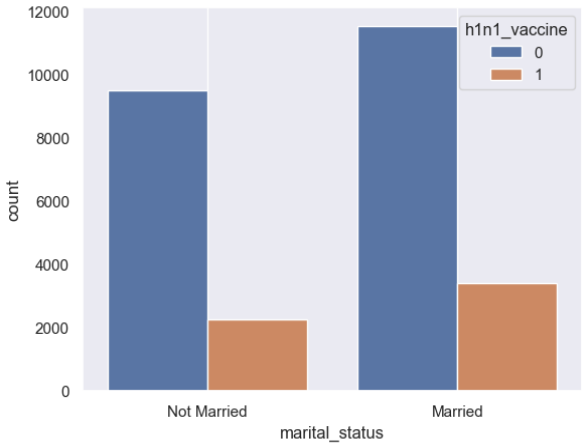
From the graph its shown that older people of age 65+ years and people having qualification college graduate are taking the h1n1 vaccine more than others.



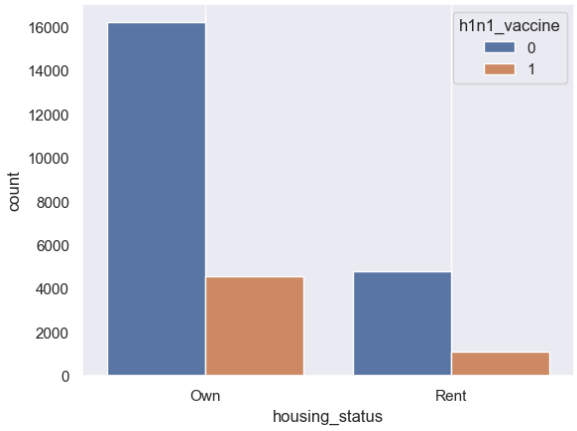


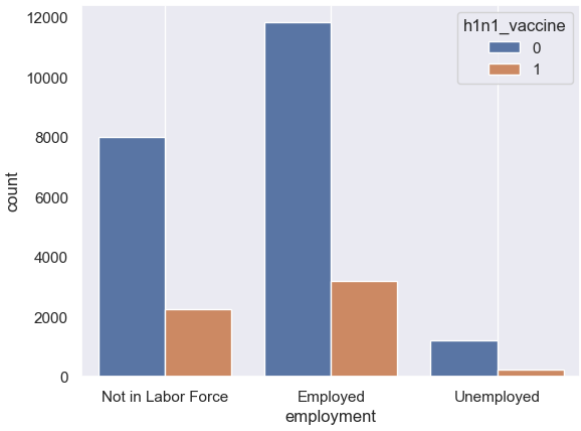
Here I observed that white race people and more of females are taking the h1n1 vaccines more.



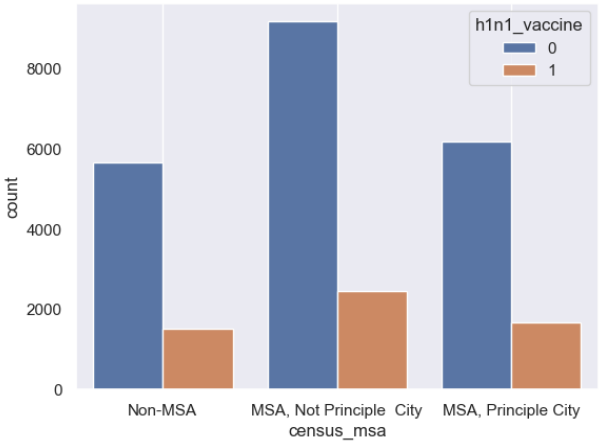


Most of married people and people having their income level more than 75,000 are taking the h1n1 vaccines.



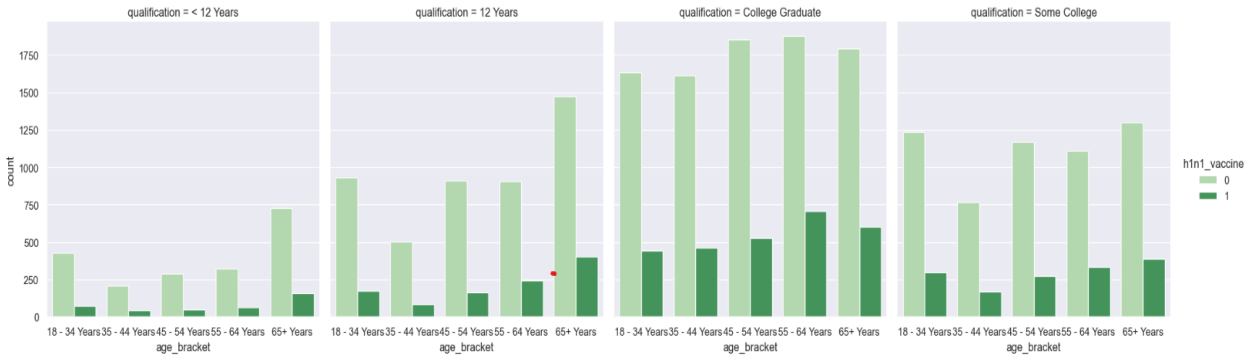


From the above graph it is obvious that peoples who took the h1n1 vaccines are more likely to have their own house and emplyoyed also.



More MSA, Not principle city peoples have taken the h1n1 vaccines.

Since our age group starts at age 18, we can logically pair with a level of qualification. Can we find any patterns among age groups and qualification who were vaccinated?

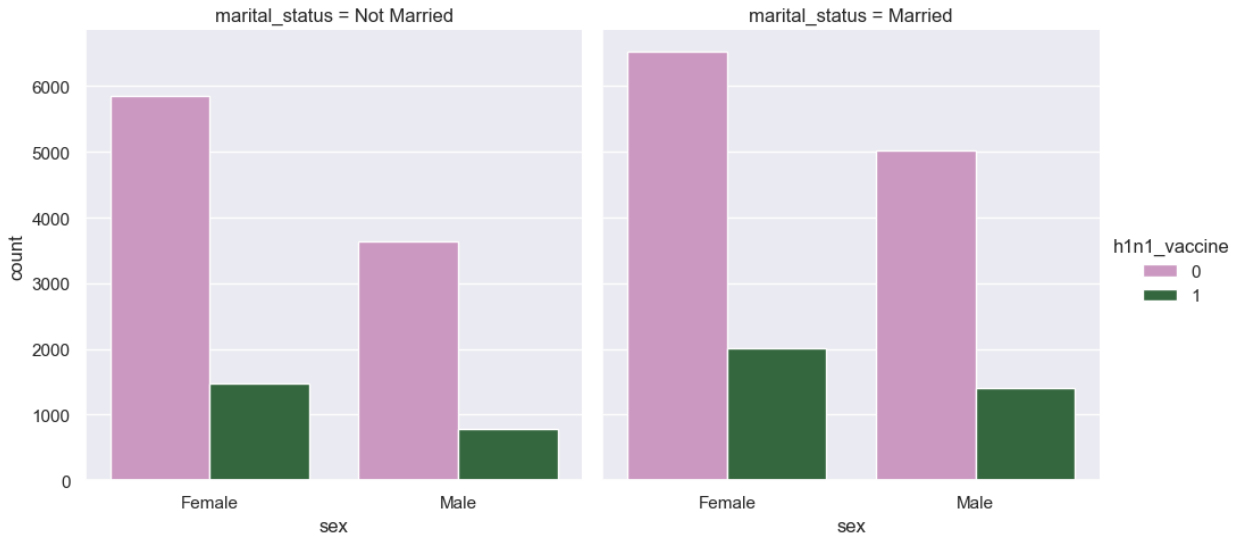


We can see generally there are more people vaccinated who are in a college graduate category. Though there isn't necessarily a blatant pattern for age groups, but it seems like more than 60years people have taken the vaccine.

Seeing the income level and housing status of people who have taken h1n1 vaccine :

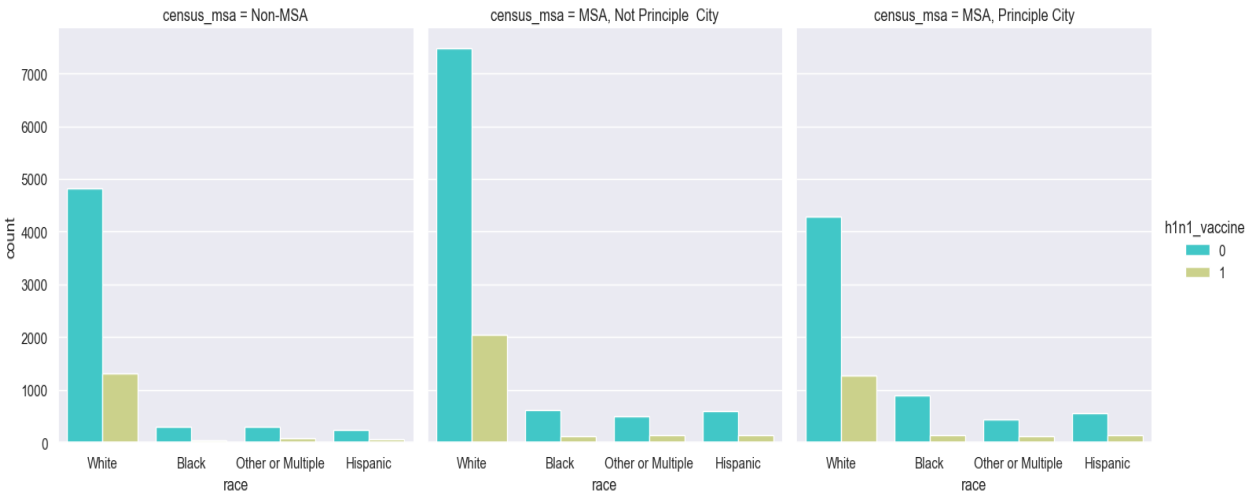
The higher the income and those who own a residence, the more someone in general received the vaccine.

Seeing gender and marital status:



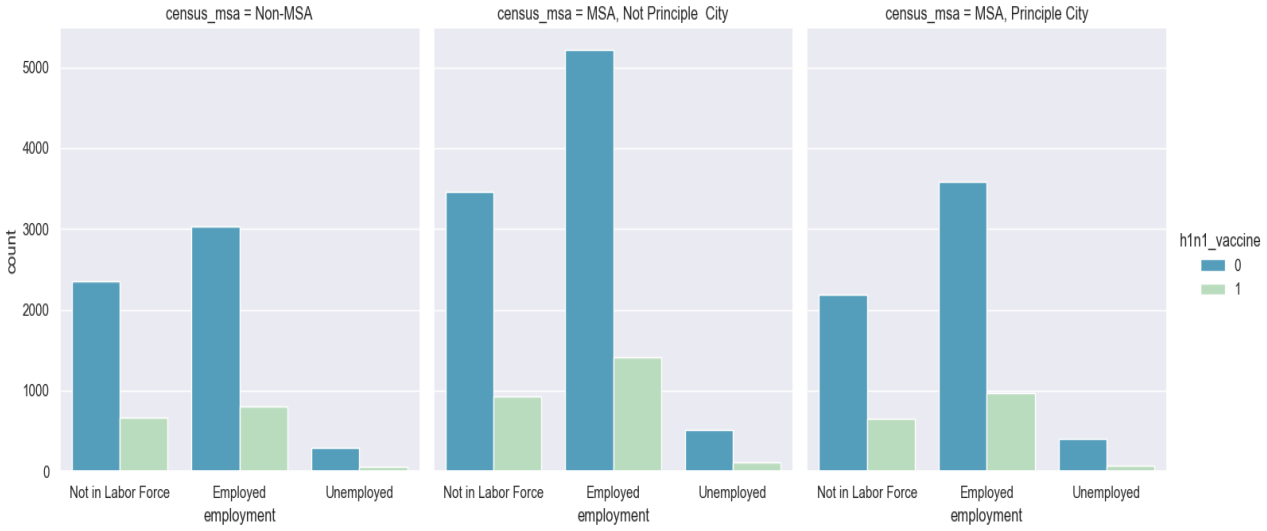
From graph we observe that married female tends to get higher vaccinations. But the number is higher for seasonal vaccine than h1n1 vaccine.

Seeing Race and census msa :



White people have taken the h1n1 vaccine of census msa=MSA, not principal city

Seeing employment and census_msa:

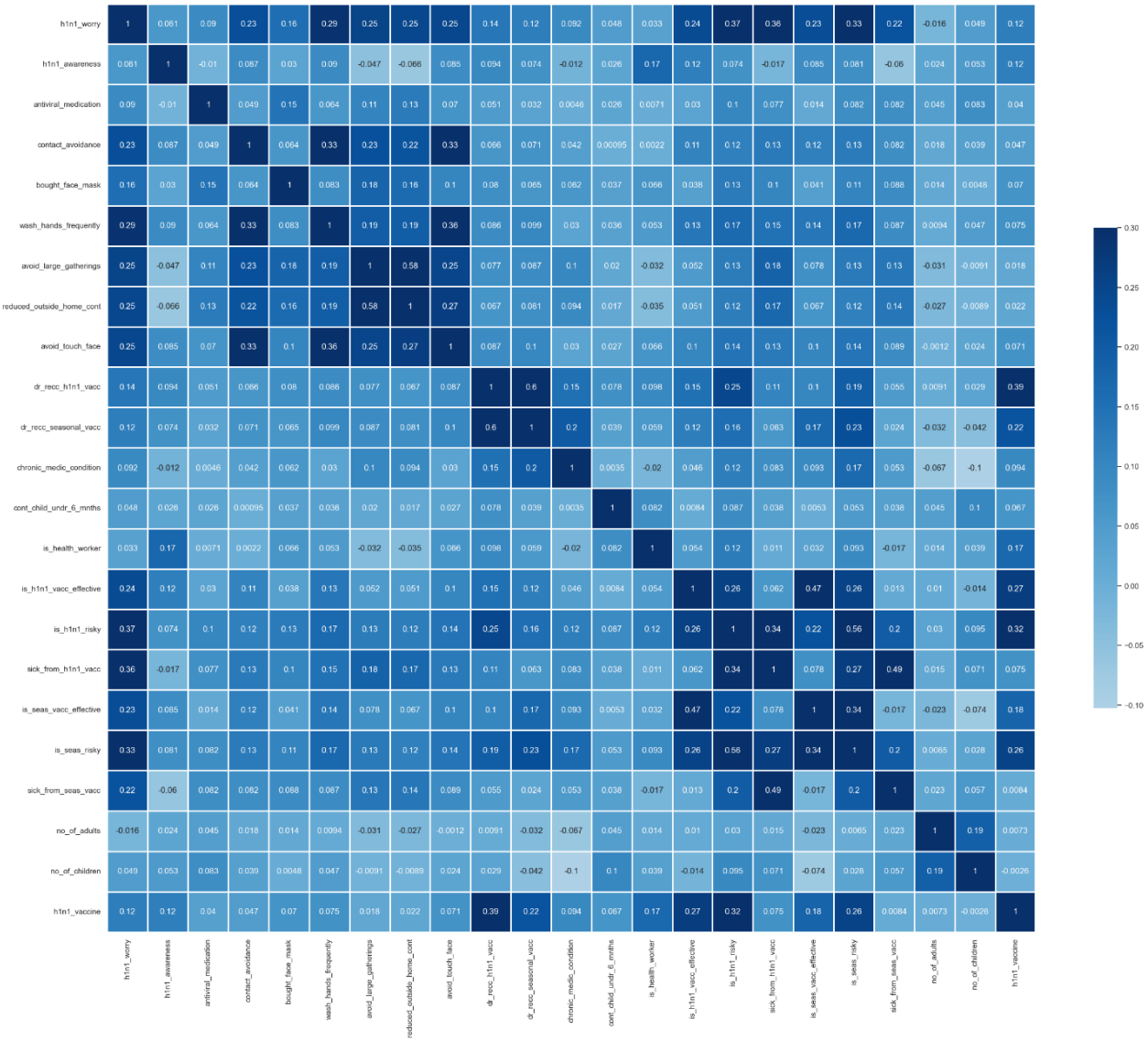


Most employed people of census_msa=MSA, not principal city have taken the vaccine.

From the above relationship I observe that "White race married female people belonging to age group 55-65yrs having their income above poverty line and their own house staying in MSA, not principal city" tends to take more vaccines than others.

Multivariate Analysis:

Correlation using Heat Map:



High Positive correlations between the 'behavioral_features' is noted, some of them may be redundant. It might be a case of multicollinearity.

High Positive Correlations between opinion of h1n1 risk, doctor recommendation of vaccines Vs whether the person really took the vaccine seems fairly obvious. Overall, the data features seem to be positively correlating with the act of taking the vaccination, except with some rare differences

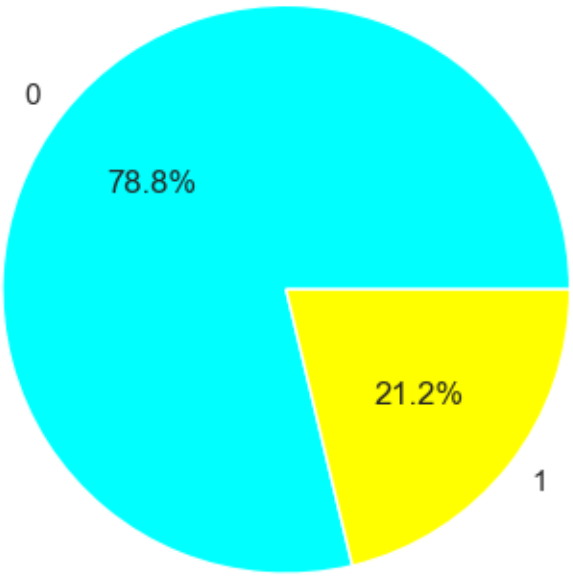
Can mainly notice that there are many redundant features/sparsely correlated features which should be taken care of.

Class Imbalance Check:

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 to get h1n1 vaccine, could be poor. From a business perspective it is important to identify or predict the people who are likely to get h1n1 vaccine and formulate strategies to minimize the h1n1 vaccine ratio.

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.

Need for variable transformation:

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   h1n1_worry                   26707 non-null  float64
 1   h1n1_awareness               26707 non-null  float64
 2   antiviral_medication         26707 non-null  float64
 3   contact_avoidance            26707 non-null  float64
 4   bought_face_mask             26707 non-null  float64
 5   wash_hands_frequently        26707 non-null  float64
 6   avoid_large_gatherings       26707 non-null  float64
 7   reduced_outside_home_cont    26707 non-null  float64
 8   avoid_touch_face             26707 non-null  float64
 9   dr_recc_h1n1_vacc            26707 non-null  float64
 10  dr_recc_seasonal_vacc        26707 non-null  float64
 11  chronic_medic_condition      26707 non-null  float64
 12  cont_child_undr_6_mnths      26707 non-null  float64
 13  is_health_worker             26707 non-null  float64
 14  is_h1n1_vacc_effective       26707 non-null  float64
 15  is_h1n1_risky                26707 non-null  float64
 16  sick_from_h1n1_vacc          26707 non-null  float64
 17  is_seas_vacc_effective       26707 non-null  float64
 18  is_seas_risky                26707 non-null  float64
 19  sick_from_seas_vacc          26707 non-null  float64
 20  age_bracket                  26707 non-null  object
 21  qualification                26707 non-null  object
 22  race                         26707 non-null  object
 23  sex                          26707 non-null  object
 24  income_level                 26707 non-null  object
 25  marital_status               26707 non-null  object
 26  housing_status               26707 non-null  object
 27  employment                   26707 non-null  object
 28  census_msa                   26707 non-null  object
 29  no_of_adults                 26707 non-null  float64
 30  no_of_children               26707 non-null  float64
 31  h1n1_vaccine                 26707 non-null  int64
dtypes: float64(22), int64(1), object(9)
memory usage: 6.5+ MB
```

In this data preprocessing step, we will be converting object data types to numerical using two techniques - get dummies and label encoding. Get dummies creates binary columns for each unique value in the object columns (employment, race, census msa), whereas label encoding assigns a numerical label to each unique value in the object column (sex, marital status, housing status).

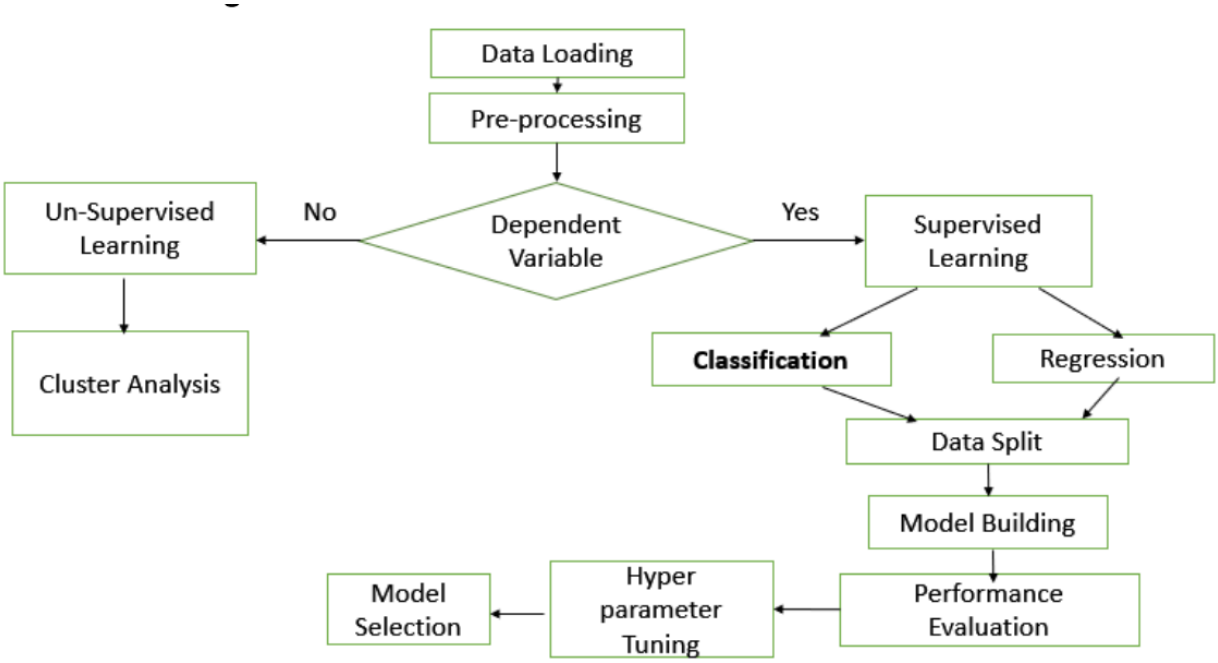
Additionally, we will be mapping the object data of (age), (qualification) and (income level) using ordinal number. This will transform the age, qualification, income level values to a common scale. Using mapping to encode ordinal numbers can make data analysis and modeling more efficient and accurate. This can simplify statistical analysis and machine learning models.

Overall, these steps will help prepare the data for machine learning models that require numerical inputs, while also ensuring that the numerical data is standardized and ready for further analysis.

	0	1	2	3	4	5
h1n1_worry	1.0	3.0	1.0	1.0	2.0	3.0
h1n1_awareness	0.0	2.0	1.0	1.0	1.0	1.0
antiviral_medication	0.0	0.0	0.0	0.0	0.0	0.0
contact_avoidance	0.0	1.0	1.0	1.0	1.0	1.0
bought_face_mask	0.0	0.0	0.0	0.0	0.0	0.0
wash_hands_frequently	0.0	1.0	0.0	1.0	1.0	1.0
avoid_large_gatherings	0.0	0.0	0.0	1.0	1.0	0.0
reduced_outside_home_cont	1.0	1.0	0.0	0.0	0.0	0.0
avoid_touch_face	1.0	1.0	0.0	0.0	1.0	1.0
dr_recc_h1n1_vacc	0.0	0.0	0.0	0.0	0.0	0.0
dr_recc_seasonal_vacc	0.0	0.0	0.0	1.0	0.0	1.0
chronic_medic_condition	0.0	0.0	1.0	1.0	0.0	0.0
cont_child_undr_6_mnths	0.0	0.0	0.0	0.0	0.0	0.0
is_health_worker	0.0	0.0	0.0	0.0	0.0	0.0
is_h1n1_vacc_effective	3.0	5.0	3.0	3.0	3.0	5.0
is_h1n1_risky	1.0	4.0	1.0	3.0	3.0	2.0

sick_from_h1n1_vacc	2.0	4.0	1.0	5.0	2.0	1.0
is_seas_vacc_effective	2.0	4.0	4.0	5.0	3.0	5.0
is_seas_risky	1.0	2.0	1.0	4.0	1.0	4.0
sick_from_seas_vacc	2.0	4.0	2.0	1.0	4.0	4.0
age_bracket	4	2	1	5	3	5
qualification	1	2	4	2	3	2
sex	0	1	1	0	0	1
marital_status	1	1	1	1	0	0
housing_status	0	1	0	1	0	0
no_of_adults	0.0	0.0	2.0	0.0	1.0	2.0
no_of_children	0.0	0.0	0.0	0.0	0.0	3.0
h1n1_vaccine	0	0	0	0	0	0
race_Hispanic	0	0	0	0	0	0
race_Other or Multiple	0	0	0	0	0	0
race_White	1	1	1	1	1	1
employment_Not in Labor Force	1	0	0	1	0	0
employment_Unemployed	0	0	0	0	0	0
census_msa_MSA, Principle City	0	0	0	1	0	1
census_msa_Non-MSA	1	0	0	0	0	0
income_level_<=75000 Above Poverty	0	0	1	0	1	1
income_level_>75000	0	0	0	0	0	0

Model Building:



The objective of the study is to predict number of people more likely to take H1N1 vaccine. We have a dependent variable which is the H1N1 Vaccine class (Binary class variable were 0 is not taking the vaccine and 1 is taking the vaccine).

Hence, we will use Supervised learning method to build classification models. The data is split into train and test at 70:30 ratio. Several classification-based algorithms from the **Scikit learn** package.

We shall begin with a simple logistic regression model which will be our base model and build several other popular technique AUC and ROC-curve, Confusion matrix and Grid Search CV which works great for binary classification. The current dataset has an issue of class imbalance. The model may have high accuracy scores and trained to predict the majority class. But our objective is to predict the minority class. Hence it is extremely important to tune the hyper parameters to increase the performance scores of the interest class and address other issues like overfitting and underfitting.

Logistic Regression:

Logit also called as maximum entropy classifier assigns probabilities to different classes. Weights are assigned to each feature which represents how important that feature is for classification decision. To find the optimal weight the algorithm uses cross-entropy loss function. The weights are running numbers from +infinity to -infinity. Hence, to transform the value to probabilities, the weights are passed through sigmoid function.

Train accuracy: 83%

Test accuracy: 84%

Class:	Class-0: Not taking vaccine		Class-1: Taking vaccine	
Train/Test	Train	Test	Train	Test
Precision	86%	86%	69%	69%
Recall	95%	95%	**41%**	**44%**
F1 Score	90%	90%	51%	54%

The model scores are quite similar between Train and Test data. Specificity for the test data is 95% which means 95% of those people who have not taken vaccine were correctly identified by the model. 44% of people who have taken the vaccine were correctly identified by the model.

Our class of interest is the **minority class**. Hence sensitivity will be an important performance parameter. The sensitivity of class-1 is very low; therefore, we need to tune the hyperparameters of the logit model and check if the performance scores improve.

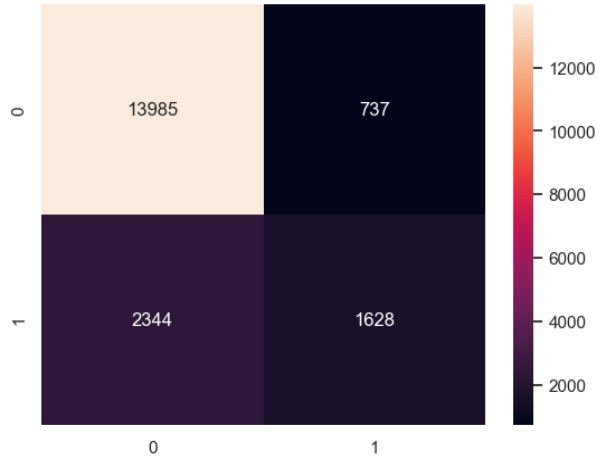
Recall is the proportion of correctly identified positive samples out of all actual positive samples in the dataset. It indicates the model's ability to identify all positive samples. In this case, the recall for both the training and test datasets is 95% for the positive class.

F1 score is the harmonic mean of precision and recall and provides a balanced measure of a model's performance. It ranges from 0 to 1, with 1 being the best possible score. In this case, the F1 score for the positive class is 90% for both the training and test datasets.

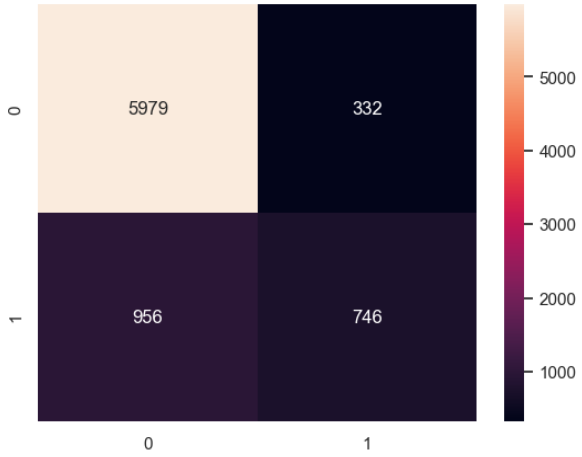
Overall, the results suggest that the model performs well in identifying the positive class, but has room for improvement in identifying the negative class. The similar performance of the model on both the training and test datasets suggests that the model is not overfitting to the training data.

Confusion Matrix:

Train data:



Test Data:

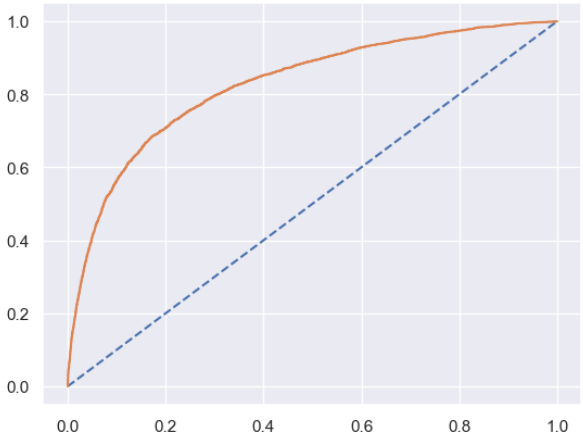


From the confusion matrix of **test data**, we can observe that there are 5979 True Negatives predicted has not taken vaccine, 746 True positives predicted has customer taken Vaccine. 332 False positives were actual is not vaccinated but predicted to be vaccinated (Type-I error) and 956 has False Negatives were actual vaccinated but predicted to be not vaccinated (Type-II error).

AUC and ROC curve:

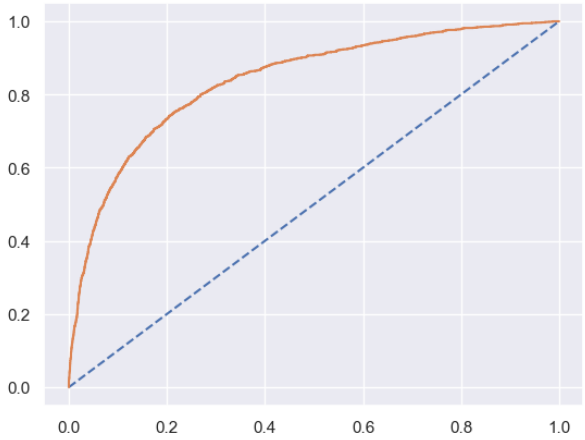
Train data:

AUC:0.82



Test Data:

AUC:0.83



The Area under curve for test is 83% which means the possibility of distinguishing between class-0 and class-1 is 83%. Receiver operating characteristic curve (ROC) is a graph between True positives and false negatives. From the graph it is visible that the curve is more inclined towards

the y axis which are true positives. (Refer Appendix for confusion matrix, AUC score and ROC curve for all the modes).

Model Performance using Grid search best params:

Grid search is a hyperparameter tuning technique that exhaustively searches through a specified set of hyperparameters to find the best combination that results in the best model performance. Once the best hyperparameters are identified, the model can be trained using those parameters to achieve the best performance.

Penalty: l2	*Solver: 'lbfgs'*	*tol: 0.01*

This set of hyperparameters configures the logistic regression model to use L2 regularization, optimize using the L-BFGS algorithm, and stop the solver when the objective function does not improve by more than 0.01.

Best Model score for train dataset is: 0.8351

Best Model score for test dataset is: 0.8392

Train:

	precision	recall	f1-score	support
0	0.86	0.95	0.90	14722
1	0.69	0.41	0.51	3972
accuracy			0.84	18694
macro avg	0.77	0.68	0.71	18694
weighted avg	0.82	0.84	0.82	18694

Test:

	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

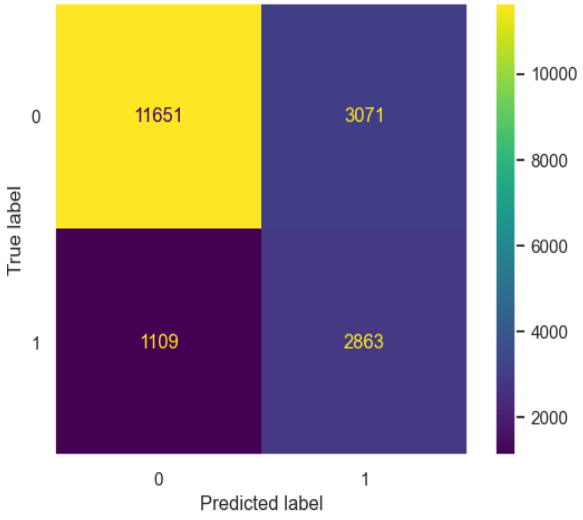
This data didn't make any changes in the confusion matrices. The sensitivity for train is 0.41 and for test it is 0.44.

Balancing confusion matrix:

Balancing a confusion matrix typically refers to adjusting the predictions of a model to improve its performance on a specific class or group of classes. Here I balance the confusion matrix by using technique **class weighting**.

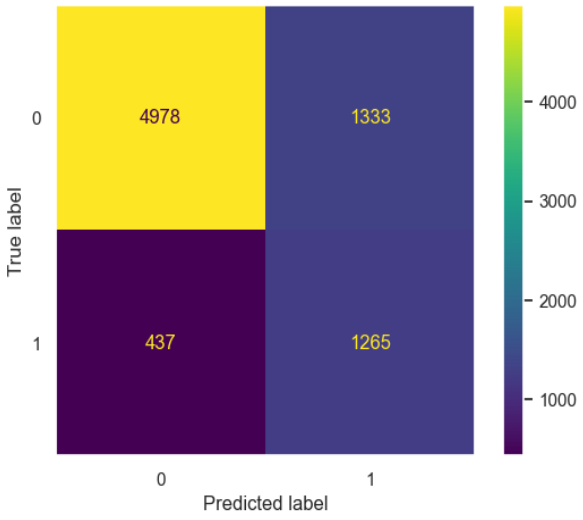
Class weighting: You can adjust the weight of each class in the model to handle class imbalance. In scikit-learn, you can do this by setting the class weight parameter to "balanced". This will automatically adjust the weight of each class based on the frequency of the class in the data.

Train data:



	precision	recall	f1-score	support
0	0.91	0.79	0.85	14722
1	0.48	0.72	0.58	3972
accuracy			0.78	18694
macro avg	0.70	0.76	0.71	18694
weighted avg	0.82	0.78	0.79	18694

Test data:



	precision	recall	f1-score	support
0	0.92	0.79	0.85	6311
1	0.49	0.74	0.59	1702
accuracy			0.78	8013
macro avg	0.70	0.77	0.72	8013
weighted avg	0.83	0.78	0.79	8013

The sensitivity for train is now 0.72 and for test it is 0.74, it means that the sensitivity (also called true positive rate) of your model has improved significantly for the minority class.

An increase in sensitivity for the test set means that the model is now better able to correctly identify the positive examples in the minority class, i.e., it has reduced the number of false negatives (i.e., cases where the model incorrectly predicts a negative outcome for a positive example). This is a desirable outcome as it indicates that the model is now performing better on the minority class and is less biased towards the majority class.

Conclusion:

In conclusion, we have successfully developed a logistic regression model to predict the likelihood of an individual receiving the H1N1 vaccine based on various demographic and health-related factors. After fine-tuning the model using grid search and hyperparameter tuning techniques, we were able to achieve a sensitivity of 74% and an accuracy of 78%, which is a significant improvement from the initial sensitivity of 44% and accuracy of 84%. The model can be used to identify individuals who are more likely to receive the H1N1 vaccine and tailor vaccination campaigns accordingly. Our findings demonstrate the importance of considering various factors when predicting vaccine uptake, and the potential value of machine learning in public health decision-making. Future research can be done to explore other predictive models and to validate the results with real-world data.

Suggestion:

Here are some suggestions to increase H1N1 vaccine uptake:

• Developing public health campaigns that focus on the benefits of getting vaccinated against H1N1, including protecting oneself and others, reducing the severity of the illness, and decreasing the risk of complications.

• Encouraging healthcare providers to recommend the vaccine to their patients, particularly those who are at a higher risk of contracting the virus or experiencing severe symptoms.

• Providing incentives for individuals to get vaccinated, such as free or discounted vaccines, and partnering with employers and schools to offer on-site vaccination clinics.

• Improving vaccine access and availability, particularly in underserved communities, by increasing the number of vaccine providers and expanding outreach efforts.

• Strengthening the vaccine supply chain to ensure an adequate supply of vaccines during outbreaks and reducing vaccine wastage through effective inventory management.

• Conducting further research to identify the reasons for vaccine hesitancy and developing targeted interventions to address these concerns.

By implementing these suggestions, we can improve H1N1 vaccine uptake and ultimately reduce the incidence of the virus and its associated healthcare costs.