PREDICTING INDIVIDUAL LIKELIHOOD OF H1N1 "SWINE FLU" VACCINATION

Project by Njiru Logan Kimathi – logankim62@gmail.com

OVERVIEW

Background Information

Beginning in spring 2009, a pandemic caused by the H1N1 influenza virus, also known as "swine flu," swept across the world. In the first year, it is estimated it was responsible for between 151,000 to 575,000 deaths globally.

A vaccine for the H1N1 flu virus became publicly available in October 2009. In late 2009 and early 2010, the United States conducted the National 2009 H1N1 Flu Survey. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, in conjunction with questions about themselves. These additional questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission. A better understanding of how these characteristics is associated with personal vaccination patterns can provide guidance for future public health efforts.

HealthFirst Alliance, the stakeholder for this project, is a non-governmental organization (NGO) dedicated to improving global public health by promoting vaccination awareness, combating vaccine hesitancy, and guiding data-driven policy decisions to enhance community health outcomes. The organization collaborates with local governments, healthcare providers, and community leaders to design and implement programs that increase vaccination rates and reduce the impact of preventable diseases.

Problem Statement

Despite the availability of vaccines for seasonal influenza and H1N1, a significant proportion of the population remains unvaccinated, exposing them to preventable health risks. Understanding the drivers of vaccine hesitancy and uptake is critical for designing effective public health strategies.

Proposed Solution: Analysis & Modelling

With a focus on H1N1 vaccination, the proposed solution involves a data-driven approach combining exploratory data analysis (EDA) and predictive modelling to uncover trends, insights, and actionable strategies for improving vaccine uptake. By leveraging machine learning models, we aim to identify the most influential factors affecting vaccine hesitancy and predict vaccine uptake for targeted public health interventions.

Steps include:

1. **Exploratory Data Analysis**: Investigate the dataset to understand demographic patterns, risk perception, health factors and existing barriers to vaccine uptake.
2. **Feature Engineering**: Use insights from EDA to preprocess data and select key variables for modelling.

3. **Baseline and Tuned Models**: Develop predictive models to assess vaccine uptake likelihood and refine them to prioritize recall, precision, and actionable outcomes.

Objectives

Main Objective

To analyze vaccine uptake patterns and develop a predictive model that identifies individuals less or more likely to receive H1N1 vaccination, enabling the design of data-driven public health campaigns to improve vaccine coverage.

Other Objectives

1. Identify key demographic, medical, and opinion-based factors influencing vaccine uptake through exploratory analysis.
2. Compare the effectiveness of different machine learning models in predicting vaccine uptake.
3. Provide actionable insights and recommendations for stakeholders to enhance vaccine awareness and uptake efforts.
4. Evaluate the relative impact of barriers such as risk perception and access to healthcare professionals in influencing vaccination behavior.

Metrics of Success

For a project aimed at **guiding public health efforts and increasing vaccine uptake**, the most important performance metrics should focus on the model's ability to identify individuals likely to vaccinate or resist vaccination, while minimizing misclassification that could misguide outreach strategies.

Hence the most important metrics of success are:

- **Recall -** the ability of the model to correctly identify individuals who are likely to vaccinate. Goal is a recall score >= 0.75.
- **Area Under Curve (AUC)** - the model's ability to distinguish between vaccinated and non-vaccinated individuals across different thresholds. Goal is an auc score >= 0.80.
- **Precision -** measures the proportion of predicted vaccinated individuals who are actually vaccinated. Goal is a precision score of >= 0.55.

DATA UNDERSTANDING

The data used for this project comes from the United States Department of Health and Human Services (DHHS), National Centre for Health Statistics.

It was collected from The National 2009 H1N1 Flu phone Survey.

The dataset contains **35 features or columns,** which can be generalized into **health and medical features, demographic and socioeconomic features, behavioral and opinion features, household and regional factors**.

For example:

- **h1n1_concern** - Level of concern about the H1N1 flu.

- **h1n1_knowledge** - Level of knowledge about H1N1 flu.
- **doctor_recc_h1n1** - H1N1 flu vaccine was recommended by doctor
- **opinion_h1n1_sick_from_vacc**- Respondent's worry of getting sick from taking H1N1 vaccine.
- **chronic_med_condition** - Has any of the following chronic medical conditions: asthma or any other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness.
- **doctor_recc_seasonal** - Seasonal flu vaccine was recommended by doctor.

It has 50,000+ records or rows, where each row in the dataset represents **one person** who responded to the survey. The entire dataset has been split into two main portions and a third portion containing just the respondent's answers on whether they received the vaccinations.

DATA PREPARATION AND ANALYSIS

This stage involved preprocessing the dataset to make it suitable for modelling later on. This included identifying and dealing with missing data or nulls, outliers or extreme values in the features and identifying and dealing with duplicate records in the dataset.

First step was combining the portions to make a full complete dataset.

Majority of the columns were of floating-point datatype i.e. have values in decimal format. The rest were of object datatype meaning their values were categorical in nature.

For missing values, the columns with the highest percentages of missing values were as follows:

health_insurance - 46%

income_poverty - 16%

employment_industry - 49%

employment_occupation - 50%

For employment_industry (Type of industry respondent is employed in) and employment_occupation (Type of occupation of respondent), their values were represented in short random character strings that were not easily understandable without extra information. I dropped these two features as result.

For health_insurance (whether respondent has health insurance), a correlation between the missing values and another feature, **income_poverty (**Household annual income of respondent with respect to 2008 Census poverty thresholds), where the income poverty is **>=$75,000(above poverty)** was discovered**.** Records matching these two conditions made up over 70% of the missing values in the health insurance column. Further investigation and domain knowledge showed that individuals above the poverty level are generally more likely to have access to health insurance, either through employment or private purchase. The null values may represent a group that is more likely to have health insurance but did not explicitly report it.

For that the missing values for these records were imputed the with the likely value of 1 - binary value in the column representing the respondent had health insurance.

income_poverty feature had categorical values so its missing values were imputed with the modal value of "<= $75,000, Above Poverty".

The resulting data frame now had very small missing value percentages across all columns and I dropped all the records with any missing values without the risk of losing a lot of data.

The data frame had no duplicate rows or any extreme outliers in any of the remaining features. This is due to the fact that its values were almost entirely grouped into “categories” even for the numerical features.

DATA ANALYSIS

This stage involved conducting exploratory data analysis (EDA) to identify trends, patterns and relationships in the data.

Univariate Analysis

First step involved analyzing the distribution of the values in some the features or columns independently. They results were as follows:

1. **h1n1_vaccine - Whether respondent received H1N1 flu vaccine.**

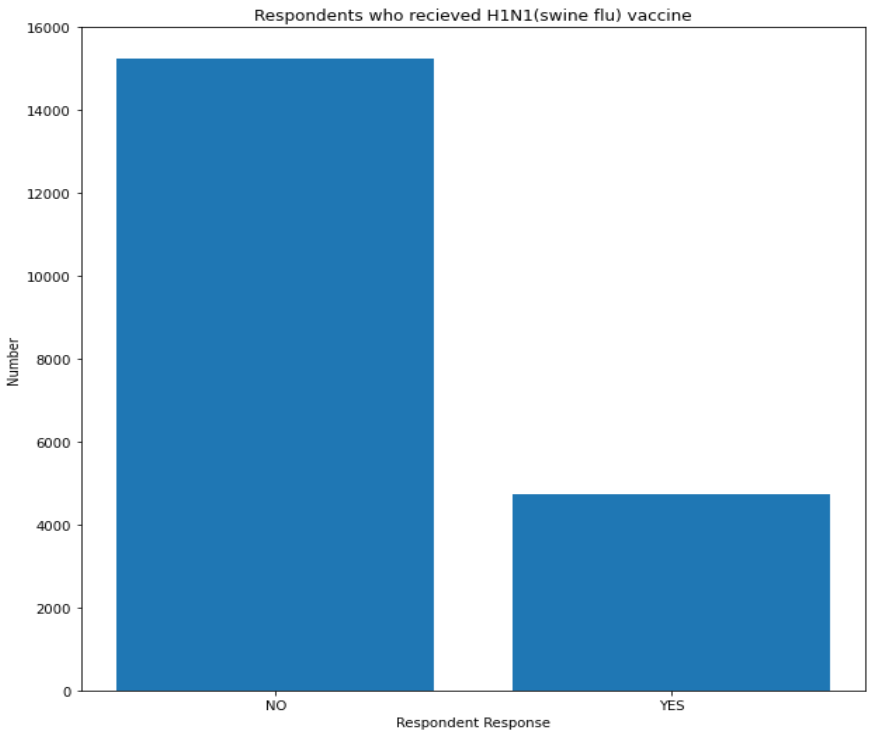

Figure 1: Individuals who received swine flu vaccination.

A majority of the respondents appear to have not received the H1N1 vaccination shots. Only 24% received the vaccination as compared to 76% who did not.

2. **Seasonal_vaccine - Whether respondent received seasonal flu vaccine.**

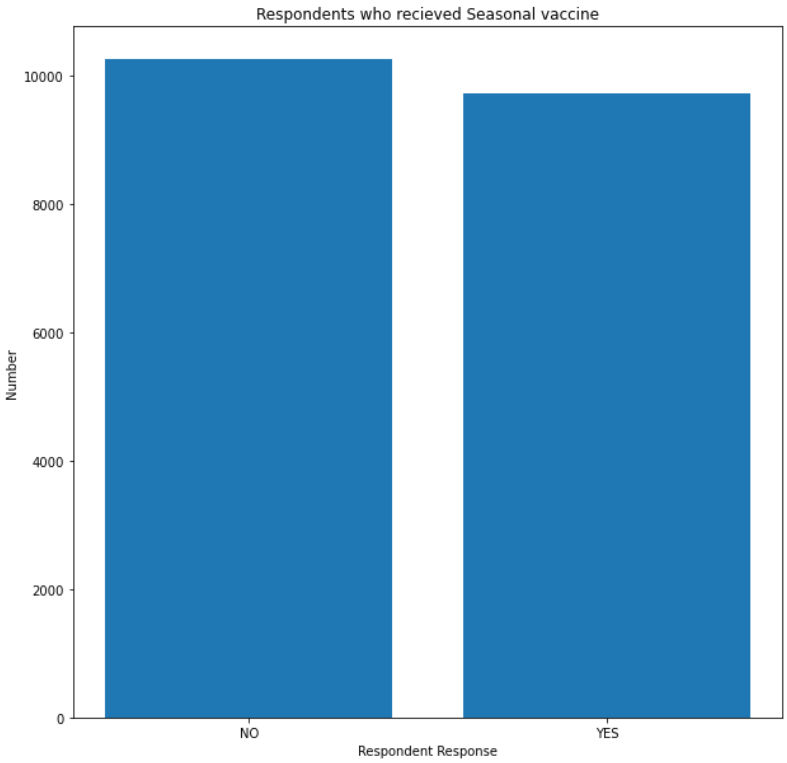

Figure 2: Individuals who received Seasonal flu vaccination.

For this one, the responses were balanced and almost equal. 51% did not receive it while 49% received the vaccination.

3. **Distribution of values across the demographic and socioeconomic features.**

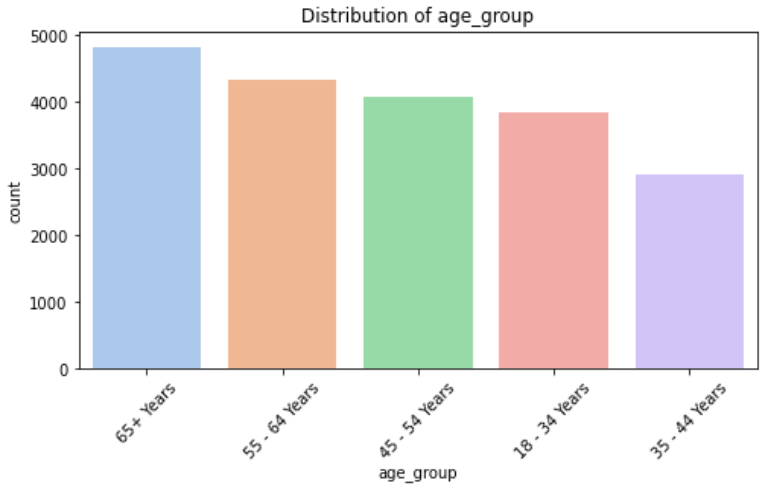

Figure 3: Age group distribution

For **Age Group**, majority of the respondents are 65+ years old (4,826 individuals) while the least represented group is 35 - 44 Years (2,919 individuals). Older individuals might have higher vaccination rates due to higher health risks.

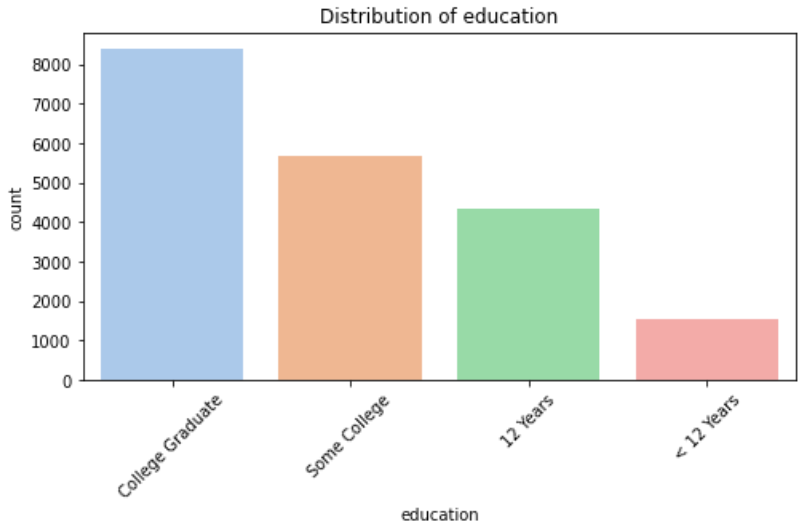

Figure 4: Education level Distribution

For **Education**, higher education levels might correlate with increased vaccine awareness as the number of respondents increased with increase in education levels.

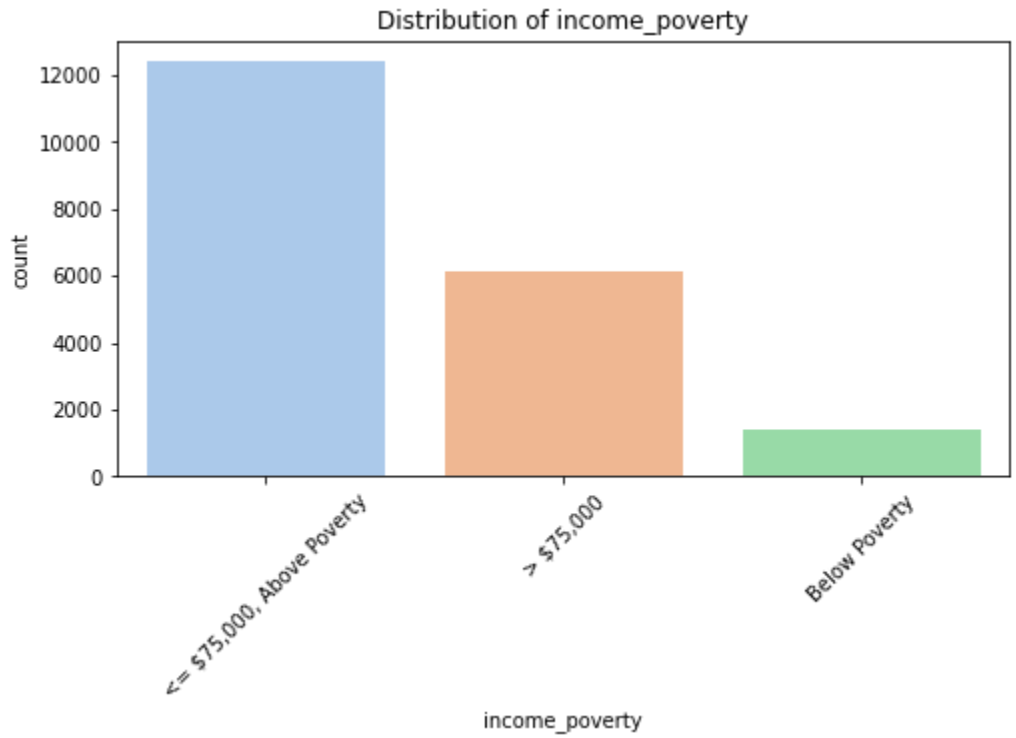

Figure 5: Annual income levels

For **Income_Poverty levels**, the majority of the respondents are **above poverty** earning upto $75,000 annually.

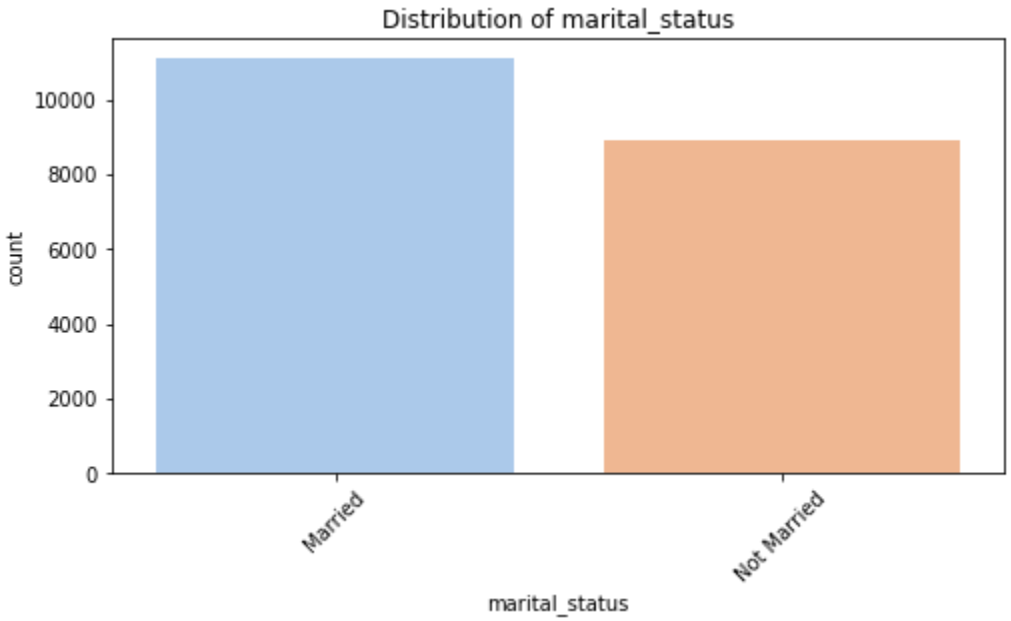

Figure 6: Marital status

Married Individuals also gave more responses than **single individuals**. This might be due to increased family responsibilities which might potentially influence vaccination rates.

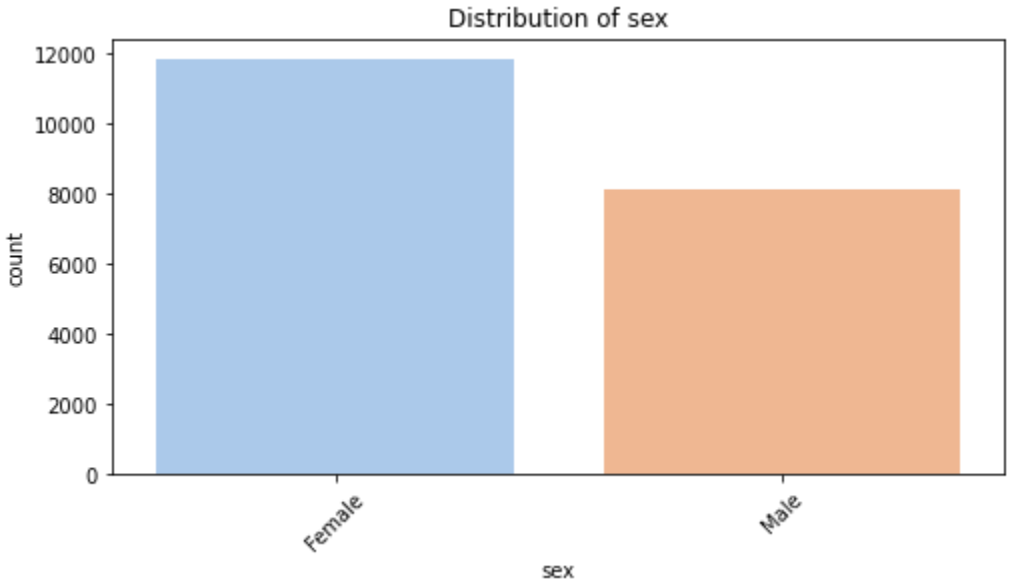

Figure 7: Gender Distribution

For **Sex**, the number of female respondents was higher than that of male respondents. 11,836 for females to 8139 for males.

4. Distribution of values across the numerical columns (with more than 2 values).

Table 1: Statistical summaries of numerical columns with more than 2 categories of values.

Feature	h1n1_concern	h1n1_knowledge	household_adults	household_children
Count	19975.000000	19975.000000	19975.000000	19975.000000
Mean	1.595344	1.305131	0.902028	0.524055
Std	0.886809	0.596594	0.746182	0.916407
Min	0.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	0.000000	0.000000
50% (Median)	2.000000	1.000000	1.000000	0.000000
75%	2.000000	2.000000	1.000000	1.000000
Max	3.000000	2.000000	3.000000	3.000000

h1n1_concern feature describes the *level of concern about the H1N1 flu of the respondent*. The values in the feature represent [0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned]

Mean: 1.59 indicates that, on average, respondents are **somewhat concerned** about H1N1 flu.

Quartiles: 25% and 75% values suggest most respondents are **between "Not very concerned" (1) and "Very concerned" (2)**.

Indicating that a significant portion of the population is **moderately or highly concerned** about H1N1, potentially indicating a readiness to accept preventive measures like vaccination.

h1n1_knowledge feature describes the *level of knowledge about H1N1 flu.* The values in the feature represent [0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.]

Mean: 1.31 indicates that respondents generally have **"A little knowledge"** about H1N1 flu.

Quartiles: Median knowledge level is 1 ("A little knowledge").

From this more can be done to educate the public about H1N1 flu.

household_children feature describes the *number of children in household, top-coded to 3.*

Mean: 0.52 indicates that many households have no children or just one child.

Quartiles: Half of the households have no children.

Households with children might be **more inclined** to vaccinate.

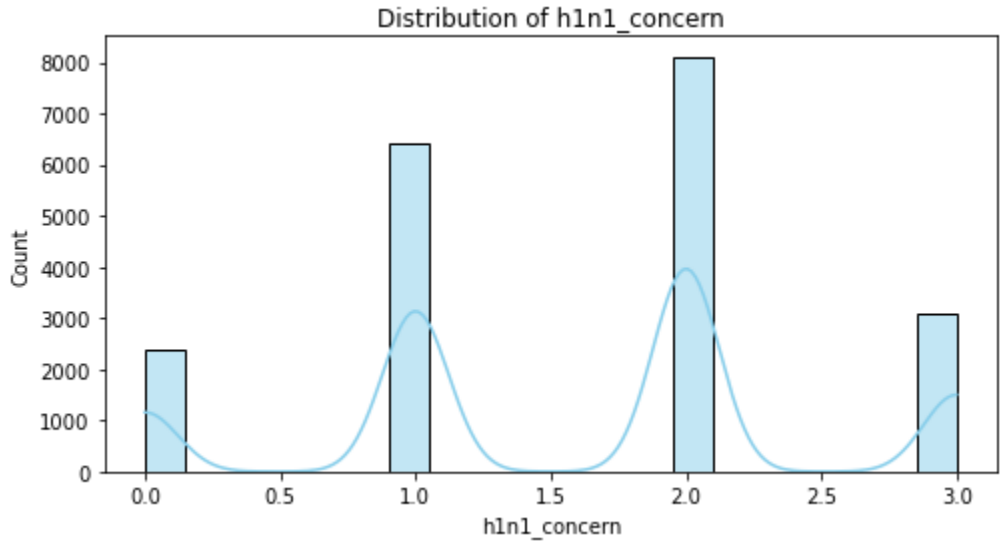

Figure 8: Histogram example of one of the numerical columns (h1n1_concern).

5. **Distribution of values across numerical features with binary values.**

A lot of the features in the dataset are made up of binary values: 0=No and 1=Yes. This analysis yielded interesting results.

Behavioral

Behavioral Antiviral Meds - Only 4.7% of respondents use antiviral medications, indicating a low adoption rate.

Behavioral Avoidance - A significant 73.7% practice avoidance behaviors to prevent flu spread, showing widespread concern and proactive measures.

Behavioral Face Mask - Only 6.5% use face masks, suggesting a potential area for improvement.

Behavioral Wash Hands - High adoption (83.1%) of handwashing indicates a strong awareness of this basic preventive measure.

Behavioral Large Gatherings - 34.5% have reduced time at large gatherings, indicating moderate adherence to social distancing.

Behavioral Outside Home - 32.2% restrict outdoor activities, suggesting majority still engage in outdoor activities despite flu concerns.

Behavioral Touch Face - 68.3% touch their faces frequently, a behavior that could increase infection risk.

Doctor Recommendations

Doctor Recommendation for H1N1 Vaccine - Only 22.6% report receiving a recommendation for the H1N1 vaccine, suggesting potential gaps in healthcare communication.

Doctor Recommendation for Seasonal Flu Vaccine - 33.7% receive recommendations for the seasonal flu vaccine, higher than H1N1 but still leaving room for improvement.

Health-Related

Chronic Medical Condition - 28.3% of respondents report chronic medical conditions, indicating a substantial at-risk group for severe flu outcomes.

Child Under 6 Months - Only 8.2% have a child under 6 months in the household, a critical group for targeted flu prevention.

Health Worker - 11.9% of respondents are health workers, who are crucial for vaccine advocacy and flu prevention.

Health Insurance - A majority of respondents report having health insurance, suggesting good access to healthcare services.

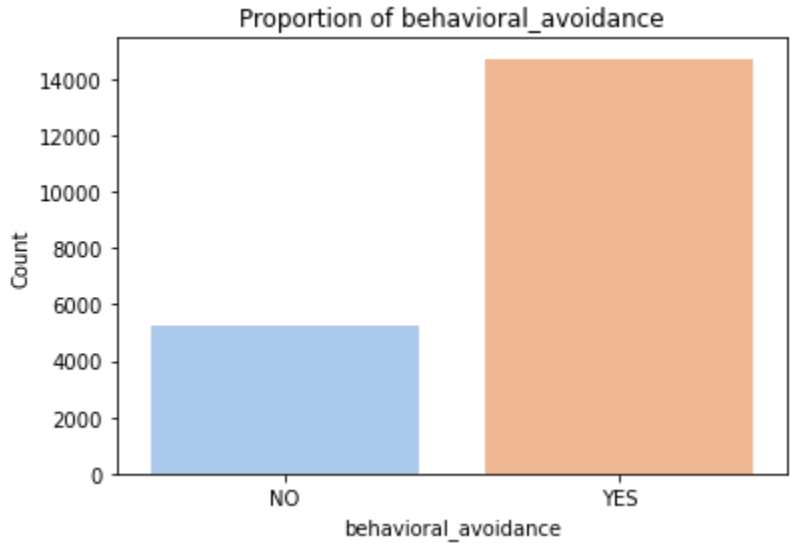

Figure 9: Number of respondents who have avoided close contact with others with flu-like symptoms

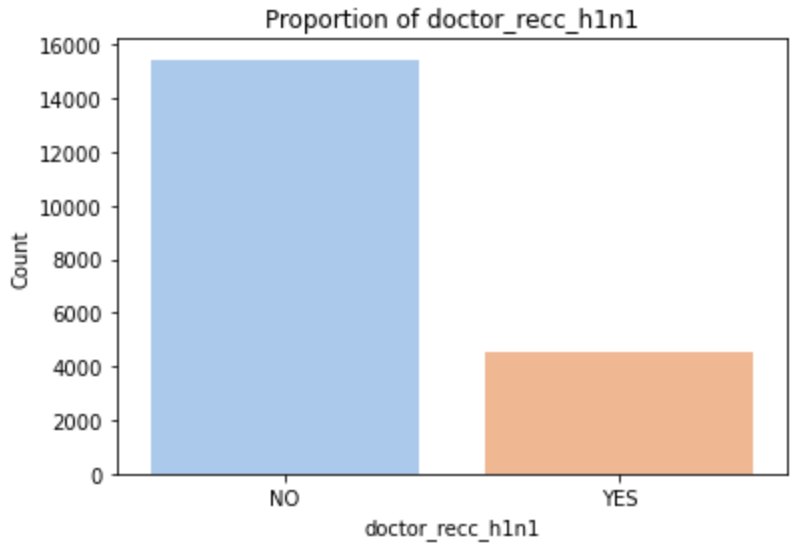

Figure 10: Number of respondents to whom H1N1 flu vaccine was recommended by a doctor

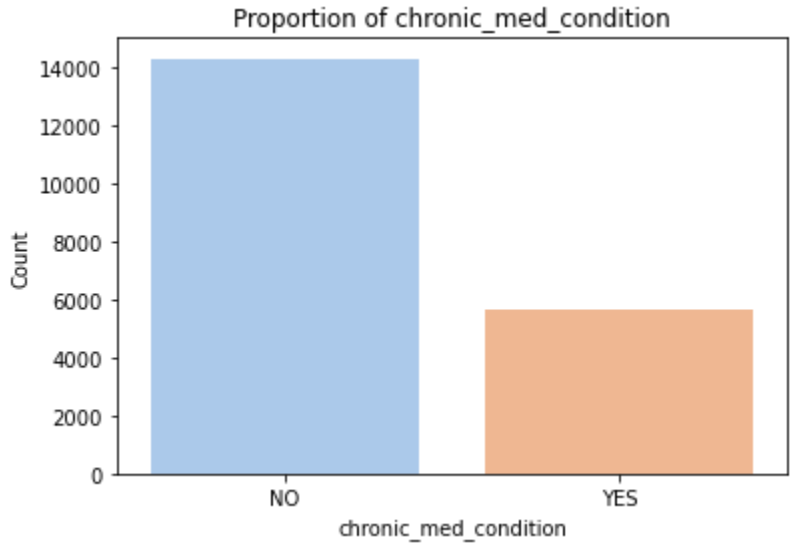

Figure 11: Number of respondents with chronic medical conditions

6. Distribution of value in opinion features.

A lot of the features are also opinion-based ordinal features. Meaning they have an inherent order ranging from "Not at all effective" to "Very effective".

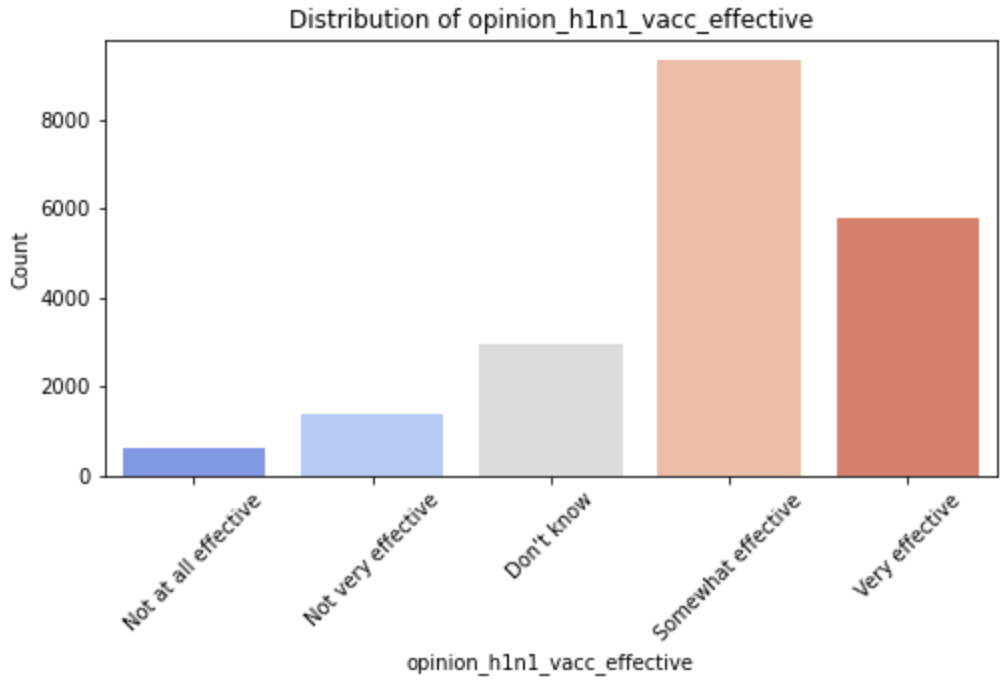

Figure 12: Respondent's opinion about H1N1 vaccine effectiveness

Most respondents believe the H1N1 vaccine is effective, with the highest scores (4.0 and 5.0) accounting for ~75% of responses

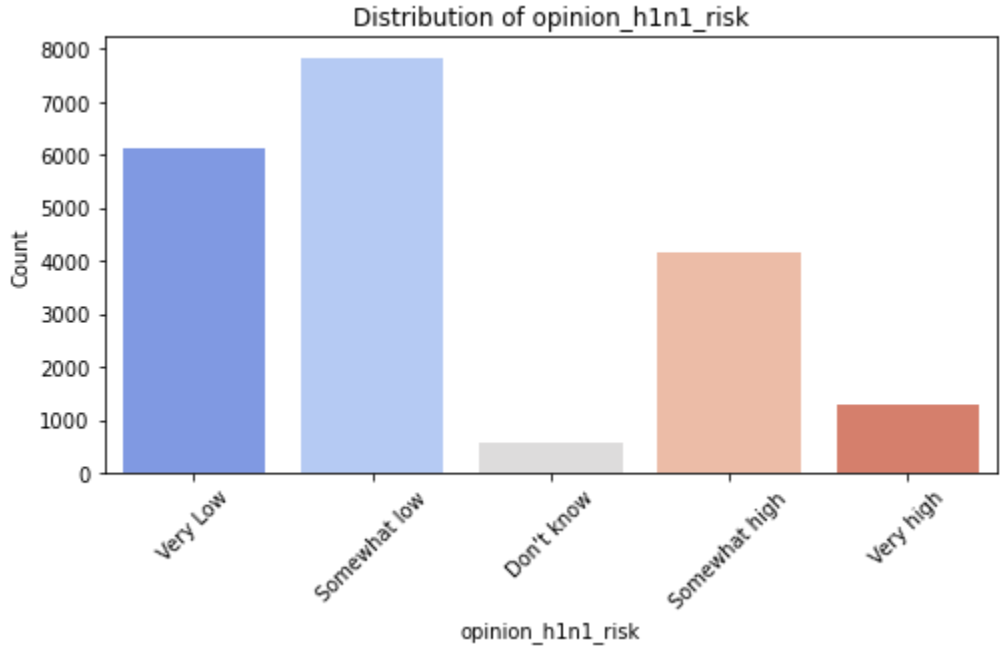

Figure 13: Respondent's opinion about risk of getting sick with H1N1 flu without vaccine

Mixed opinions: 2.0 (Moderate Risk) is the most frequent response (~39%).

1.0 (Low Risk) is the second most frequent (~31%).

A smaller group perceives a high risk (4.0 and 5.0).

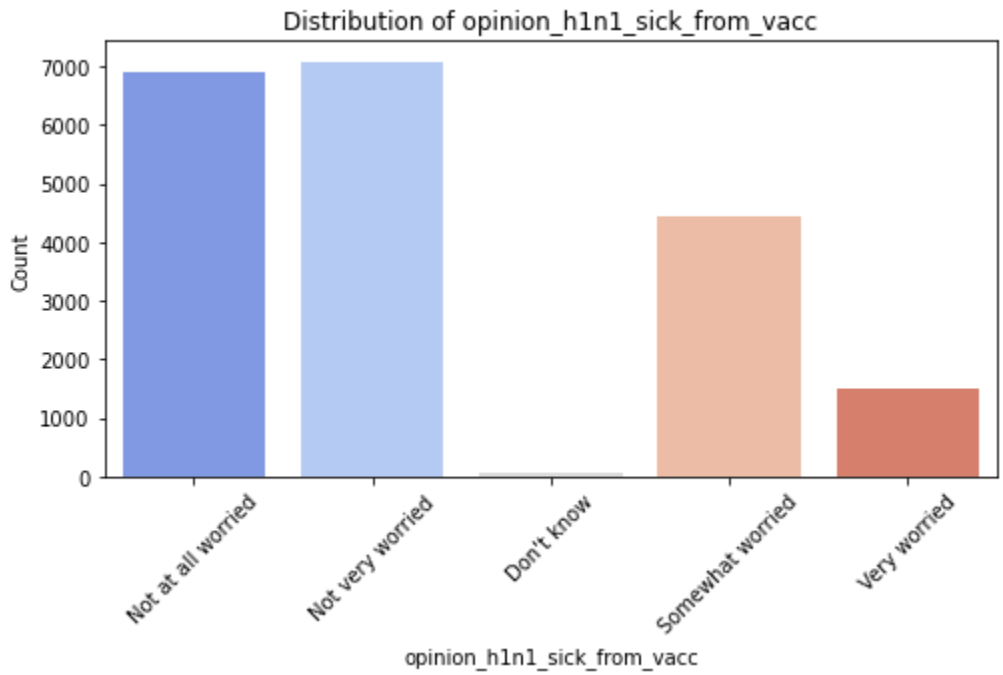

Figure 14: Respondent's worry of getting sick from taking H1N1 vaccine.

The majority of responses are 1.0 and 2.0 (Low Concern), indicating most believe the vaccine is safe.

However, some respondents (4.0 and 5.0) perceive a significant risk of sickness

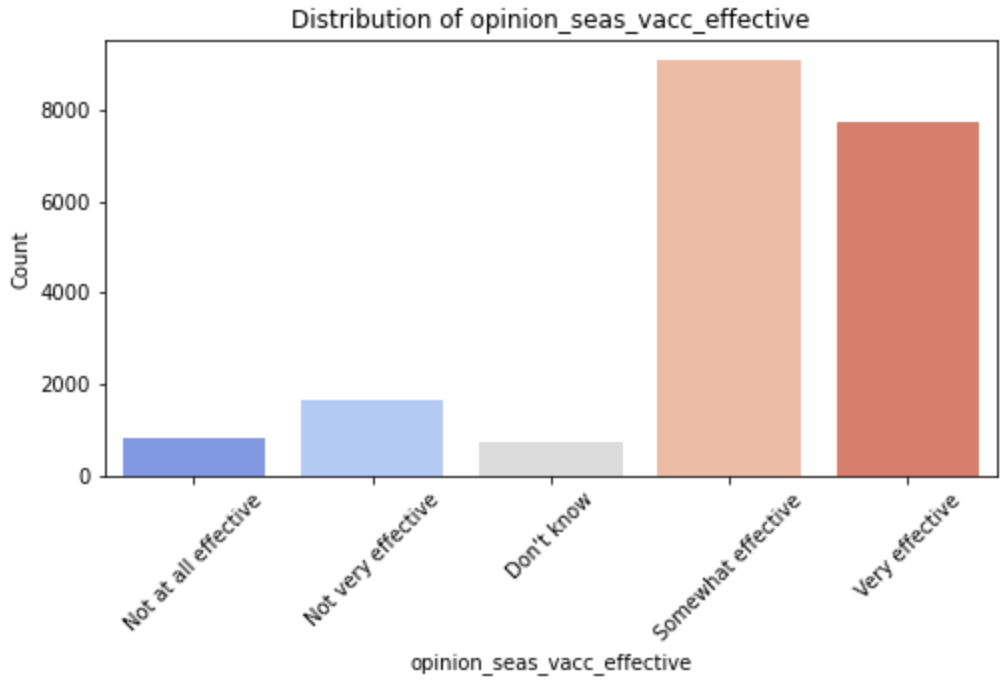

Figure 15: Respondent's opinion about seasonal flu vaccine effectiveness

Strong agreement on the effectiveness, with 4.0 and 5.0 accounting for ~85% of responses.

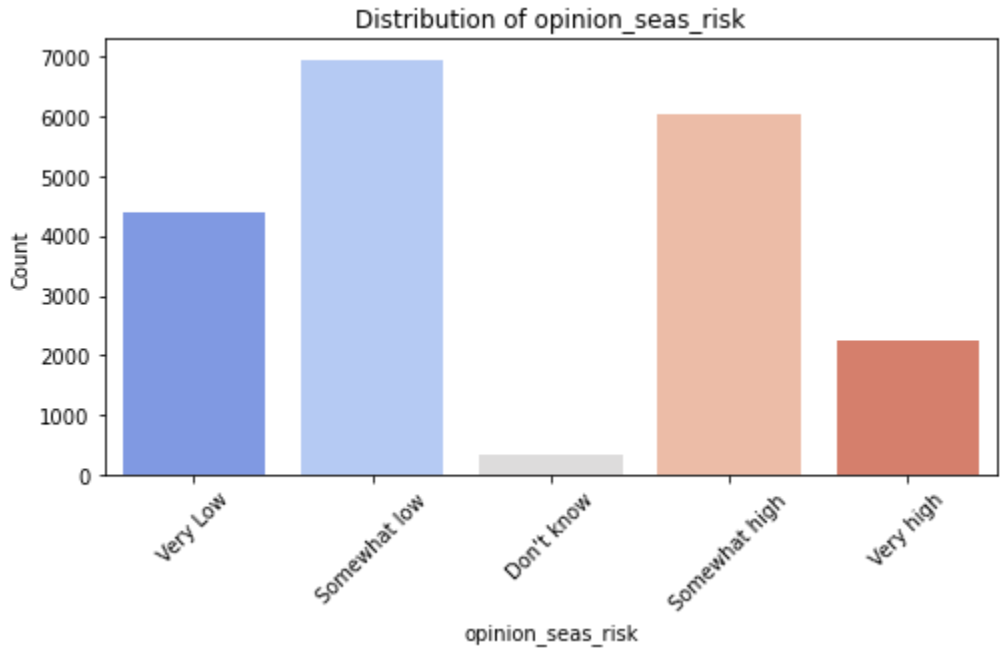

Figure 16: Respondent's opinion about risk of getting sick with seasonal flu without vaccine

2.0 (Moderate Risk) and 4.0 (High Risk) are the most common responses, indicating higher risk awareness than for H1N1

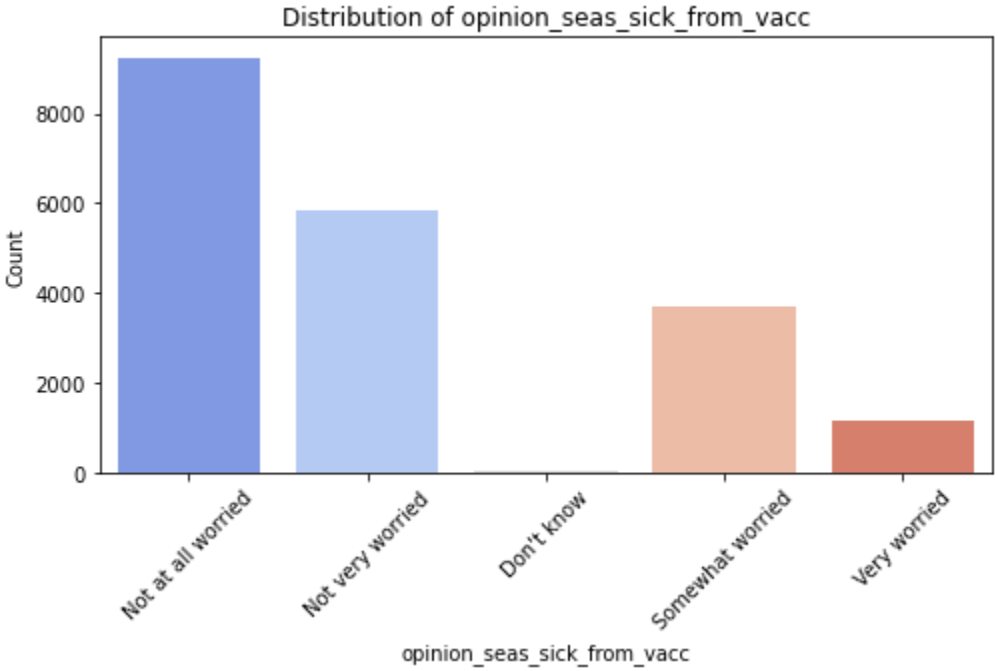

Figure 17: Respondent's worry of getting sick from taking seasonal flu vaccine

Most respondents perceive the vaccine as safe, with 1.0 (No Concern) accounting for ~46%.

Bivariate Analysis

Using **h1n1_vaccine - Whether respondent received H1N1 flu vaccine** feature as the main target, we conduct some bivariate analysis to explore the relationships between other features and our target.

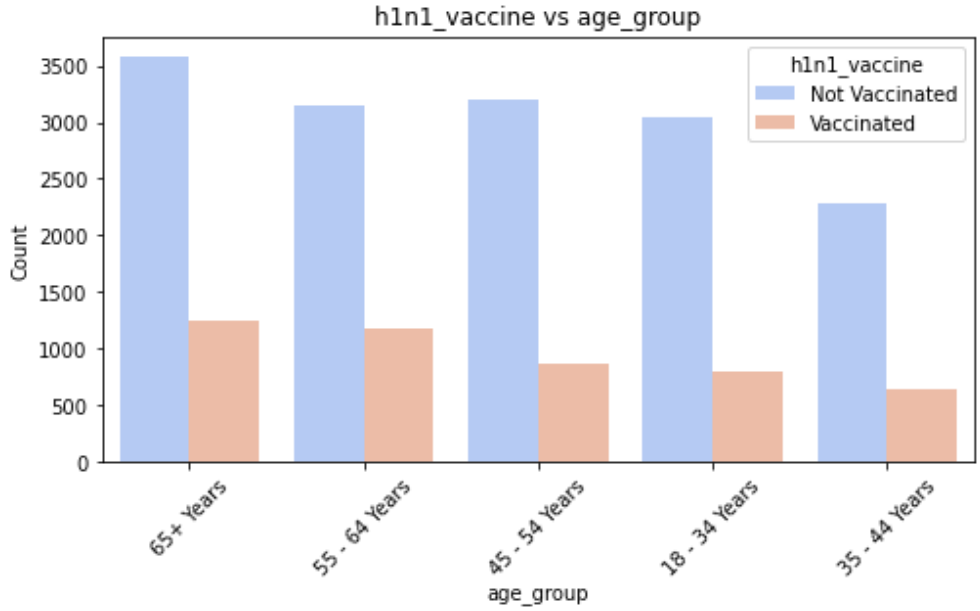

Figure 18: Age Group vs H1N1 Vaccine uptake

Age Group vs H1N1 Vaccine: *Older age groups* (55-64 years and 65+ years) have higher vaccination rates (~27% and ~26%, respectively). *Younger age groups* (18-34 years and 35-44 years) show lower vaccination rates (~21% and ~22%, respectively).

Older individuals may have a higher perceived risk or concern about the flu.

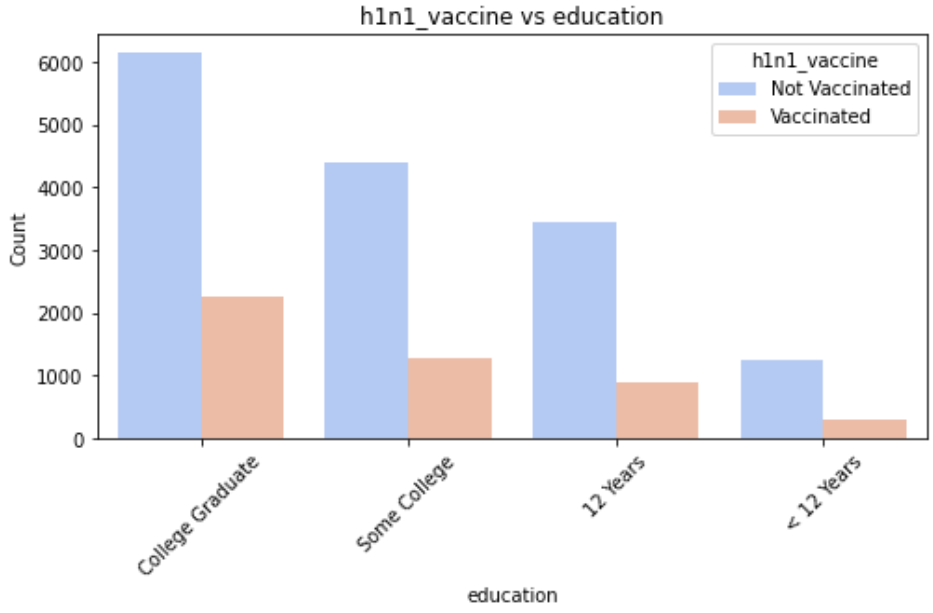

Figure 19: Education vs H1N1 Vaccine uptake

Education vs H1N1 Vaccine: *College graduates* have the highest vaccination rate (~27%), followed by individuals with some college education (~23%). Those with *less than 12 years* of education are the least likely to get vaccinated (~19%).

Higher education levels may correlate with better health awareness and access to healthcare resources

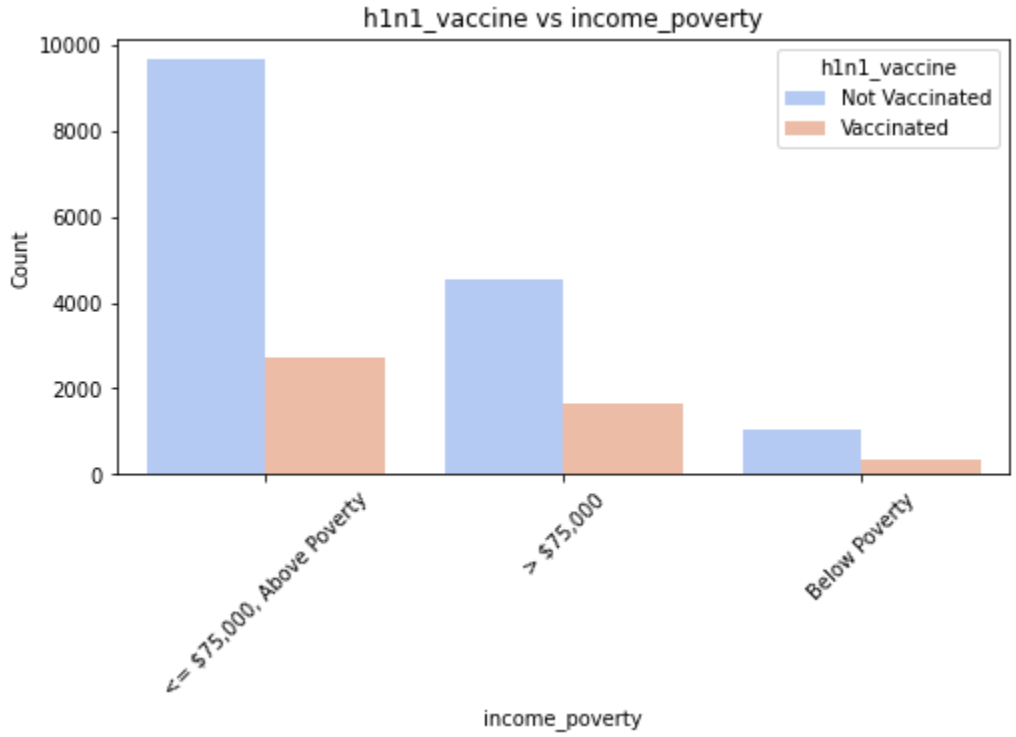

Figure 20: Annual income vs H1N1 vaccine uptake

Higher income groups (> $75,000) show the highest vaccination rate (~27%), followed by the middle-income group (<= $75,000, Above Poverty, ~22%). Those *below poverty level* have a slightly lower vaccination rate (~25%).

Income levels may influence access to vaccination or healthcare resources.

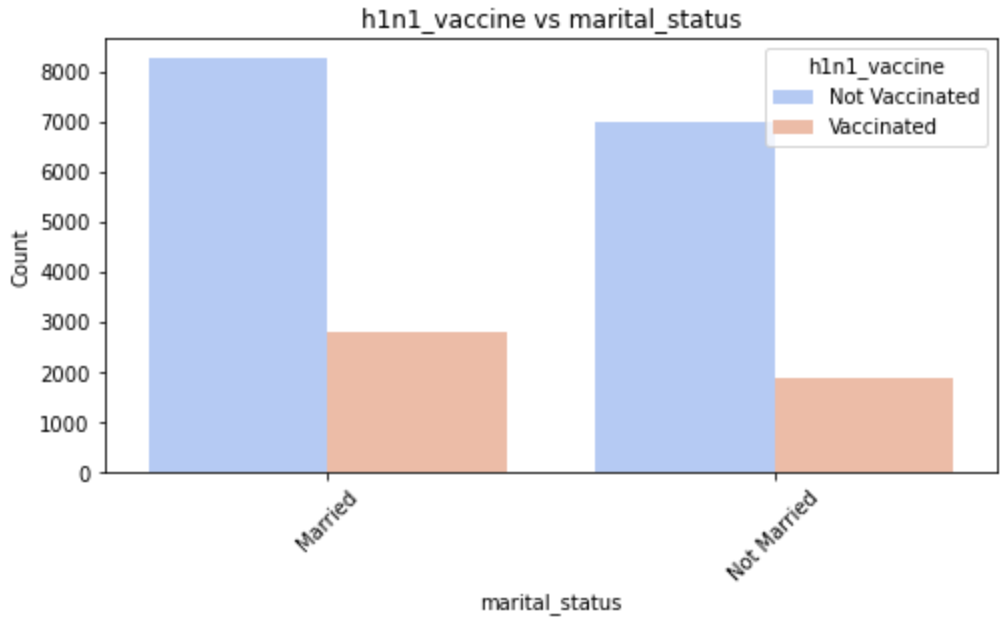

Figure 21: Marital Status vs H1N1 Vaccine uptake

Marital Status vs H1N1 Vaccine: *Married individuals* have a higher vaccination rate (~25%) compared to *non-married individuals* (~21%).

Married individuals may prioritize family health, leading to higher vaccination rates against swine flu.

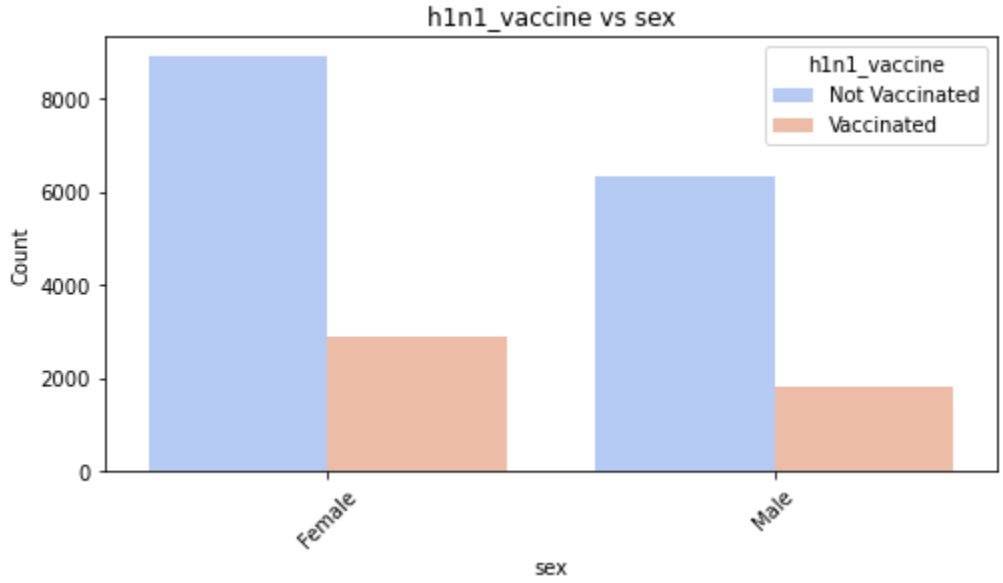

Figure 22: Gender vs H1N1 Vaccine uptake

Sex vs H1N1 Vaccine: *Females* are more likely to be vaccinated (~25%) than *males* (~22%).

Women may be more proactive in seeking preventive healthcare measures.

Table 2: Summary statistics for h1n1_concern by h1n1_vaccine

h1n1_vaccine	Count	Mean	Std	Min	25%	50%	75%	Max
0	15246.0	1.524269	0.882560	0.0	1.0	2.0	2.0	3.0
1	4729.0	1.824487	0.861412	0.0	1.0	2.0	2.0	3.0

H1N1 Concern vs H1N1 vaccine uptake: *Non-vaccinated* mean concern level: 1.52. *Vaccinated mean* concern level: 1.82.

Those who were vaccinated had a higher average level of concern about H1N1. This indicates that perceived severity may drive vaccine uptake.

Table 3: Summary statistics for h1n1_knowledge by h1n1_vaccine

h1n1_vaccine	Count	Mean	Std	Min	25%	50%	75%	Max
0	15246.0	1.26466	0.593454	0.0	1.0	1.0	2.0	2.0
1	4729.0	1.43561	0.587994	0.0	1.0	1.0	2.0	2.0

H1N1 Knowledge vs H1N1 vaccine uptake: *Non-vaccinated* mean knowledge level: 1.26. *Vaccinated* mean knowledge level: 1.43

Individuals with more knowledge about H1N1 were more likely to get vaccinated. This suggests that educational campaigns can be crucial for increasing vaccination rates.

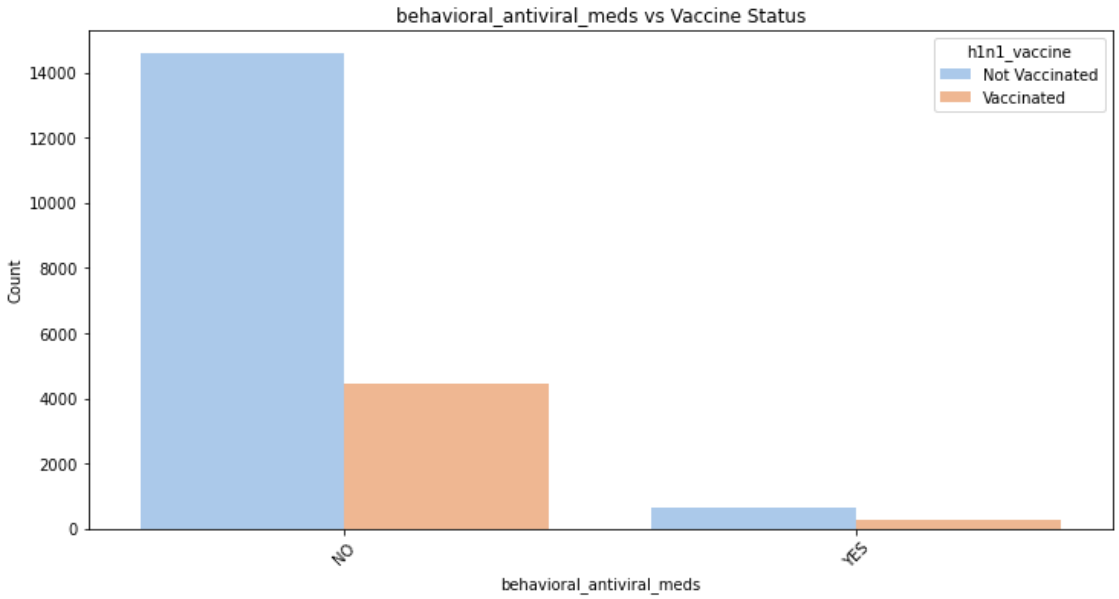

Figure 23: Antiviral use vs H1N1 vaccine uptake

Most of those vaccinated **did not use antiviral medications**. This is not surprising given only 4.7% of the respondents use antiviral medications.

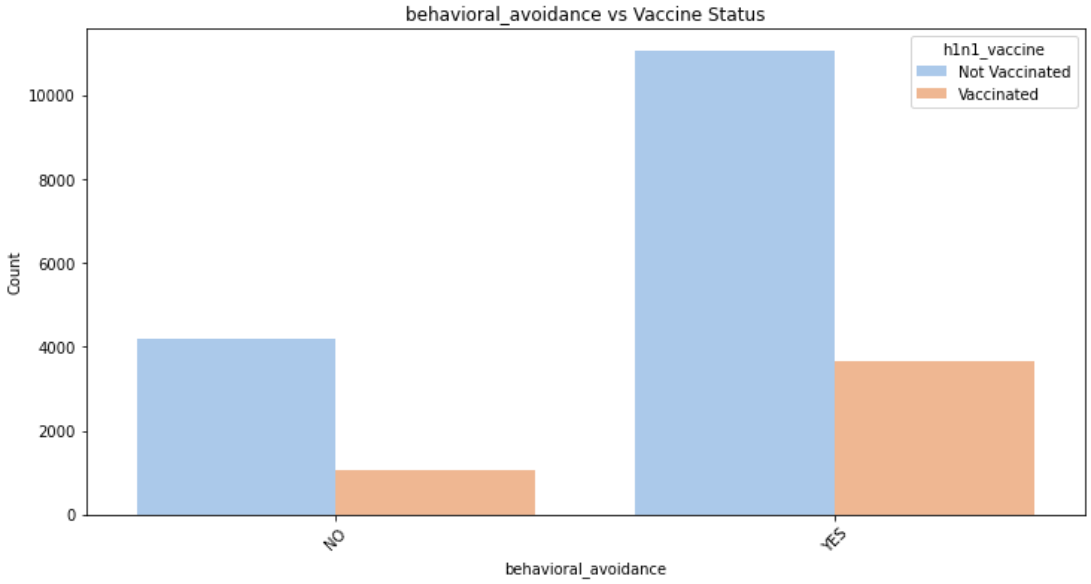

Figure 24: Avoidance behaviors vs H1N1 vaccine uptake

Majority of those vaccinated **practiced avoidance behaviors with people with flu-like symptoms**, indicating a level of caution.

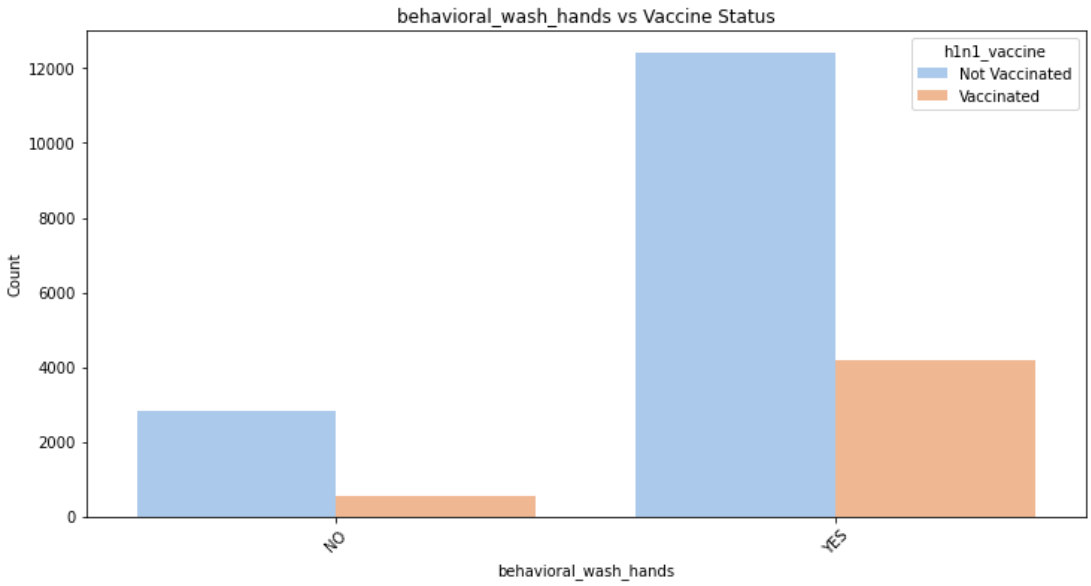

Figure 25: Handwashing behavior vs H1N1 vaccine uptake

Most of those vaccinated also **practiced handwashing**.

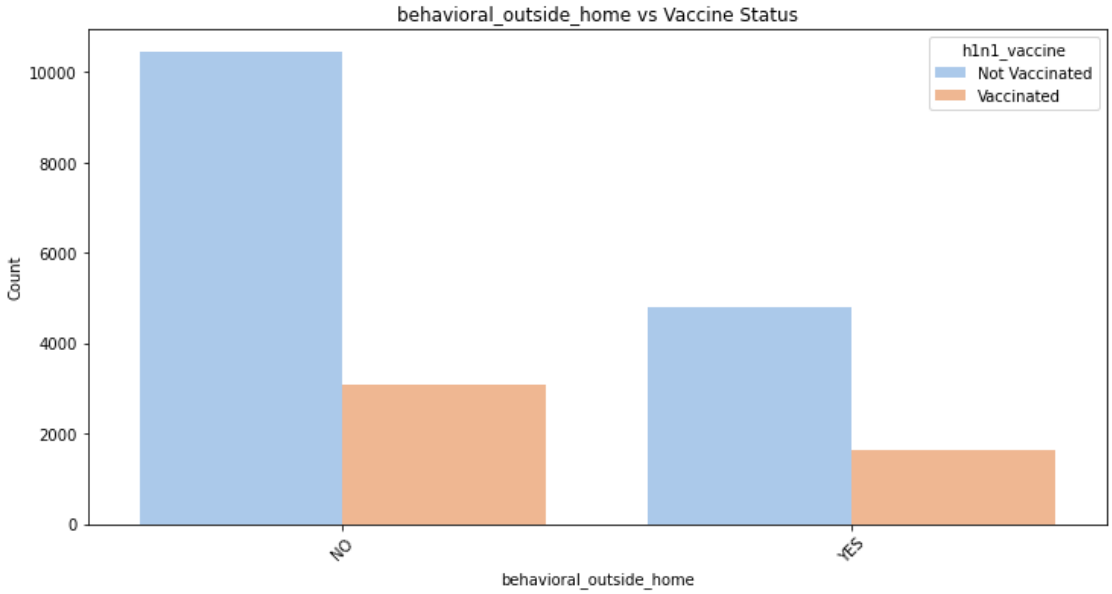

Figure 26: Reduced contact with non-family members vs H1N1 vaccine uptake

There was a **slight difference** between the number of vaccinated people who **reduced contact with people outside of their own households**. Same case for people who **reduced time spent at large gatherings**.

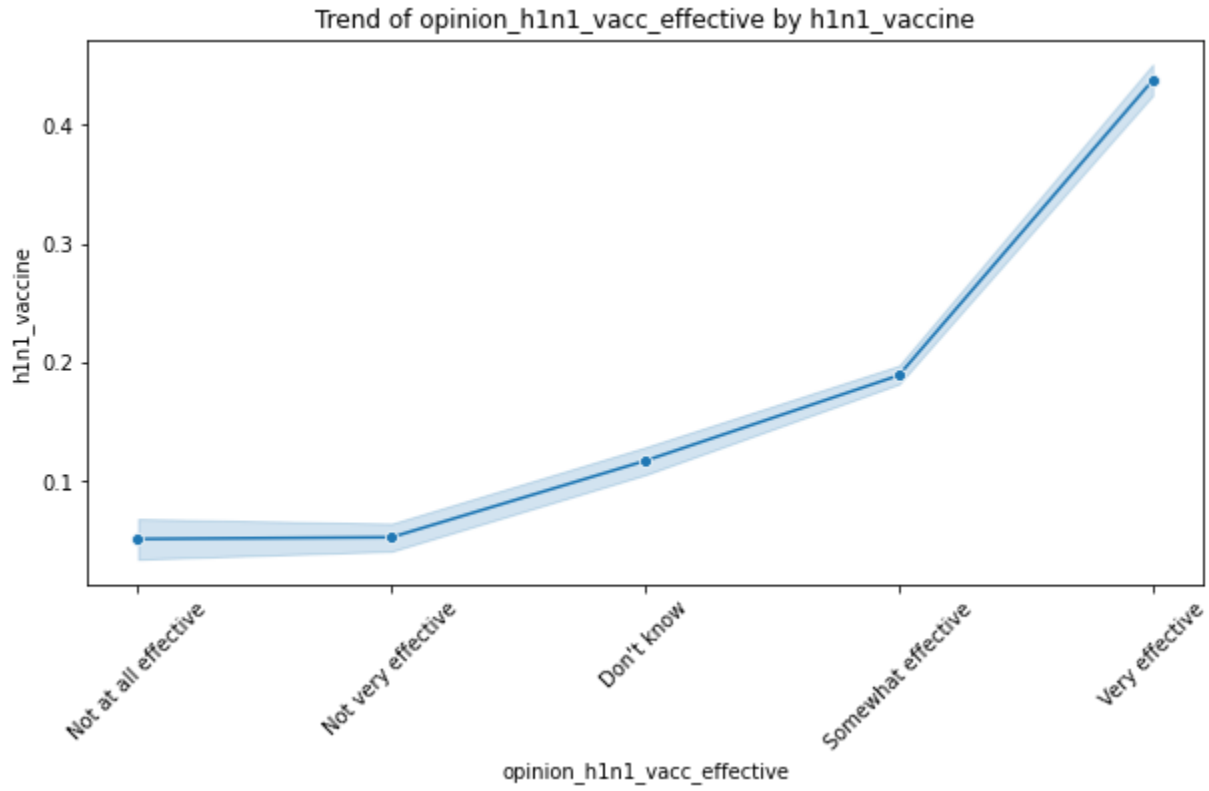

Figure 27: Opinion of H1N1 vaccine effectiveness vs H1N1 vaccine uptake

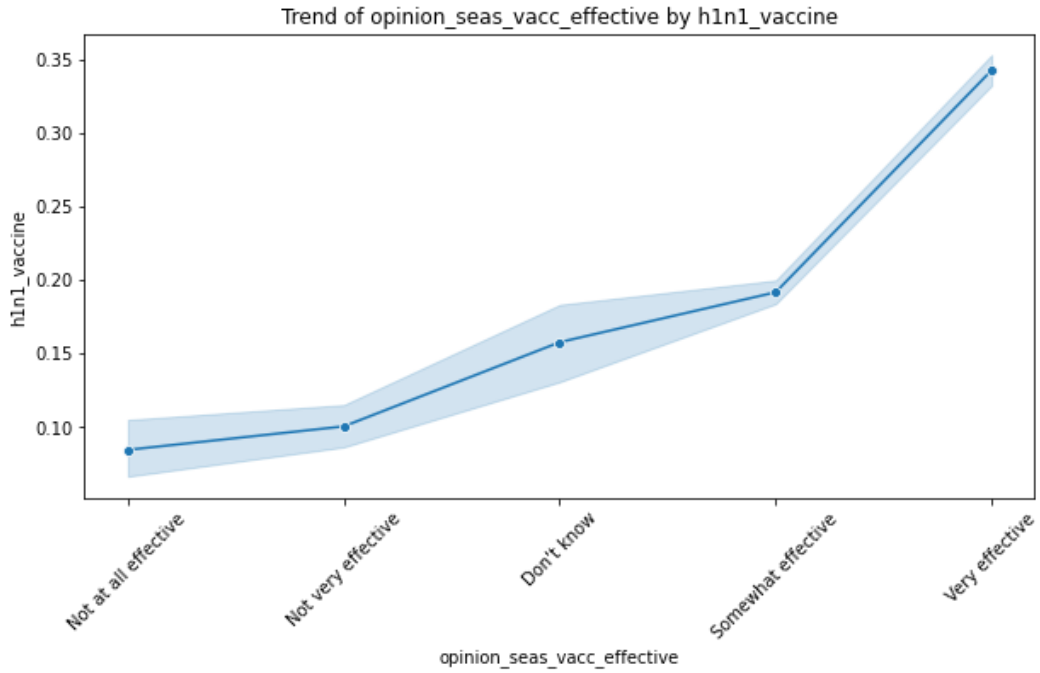

Figure 28: Opinion of Seasonal flu vaccine effectiveness vs H1N1 vaccine uptake

Unsurprisingly, vaccination uptake increased with believe in the effectiveness of the H1N1 vaccine and seasonal flu vaccines.

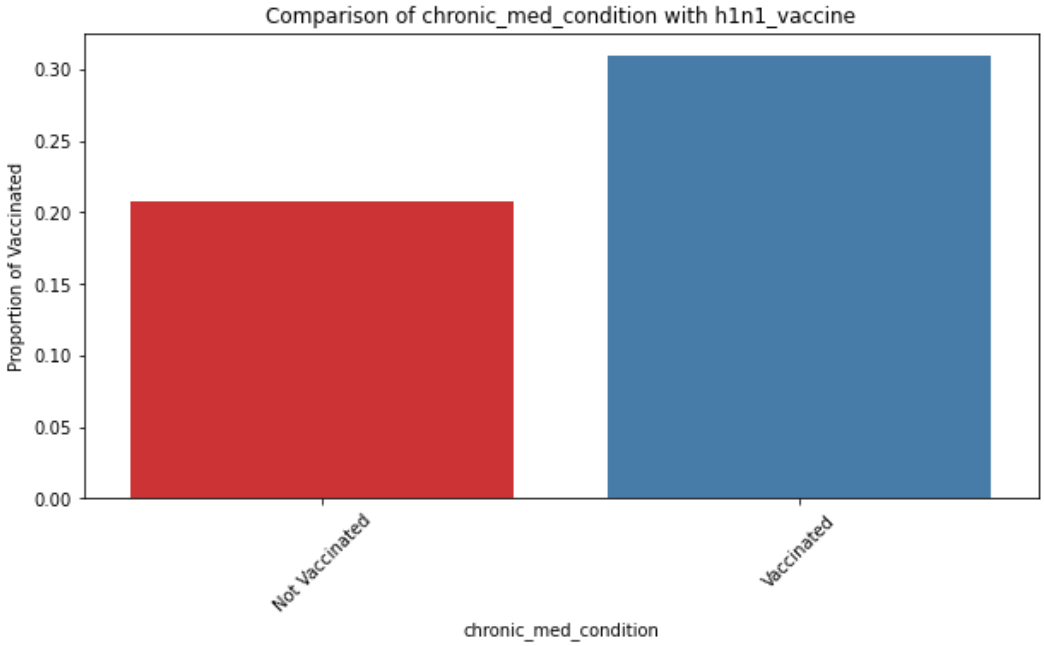

Figure 29: Chronic medical conditions vs H1N1 vaccination

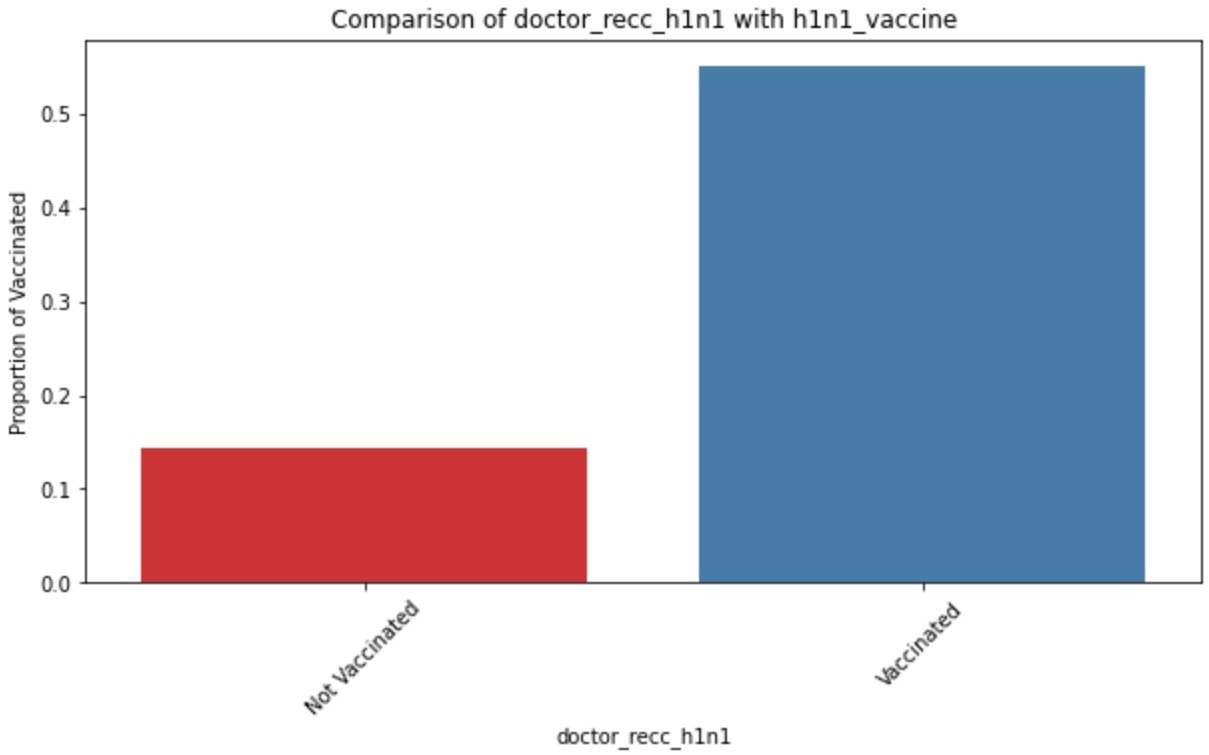

Figure 30: Doctor recommendation to get H1N1 vaccine vs H1N1 vaccination

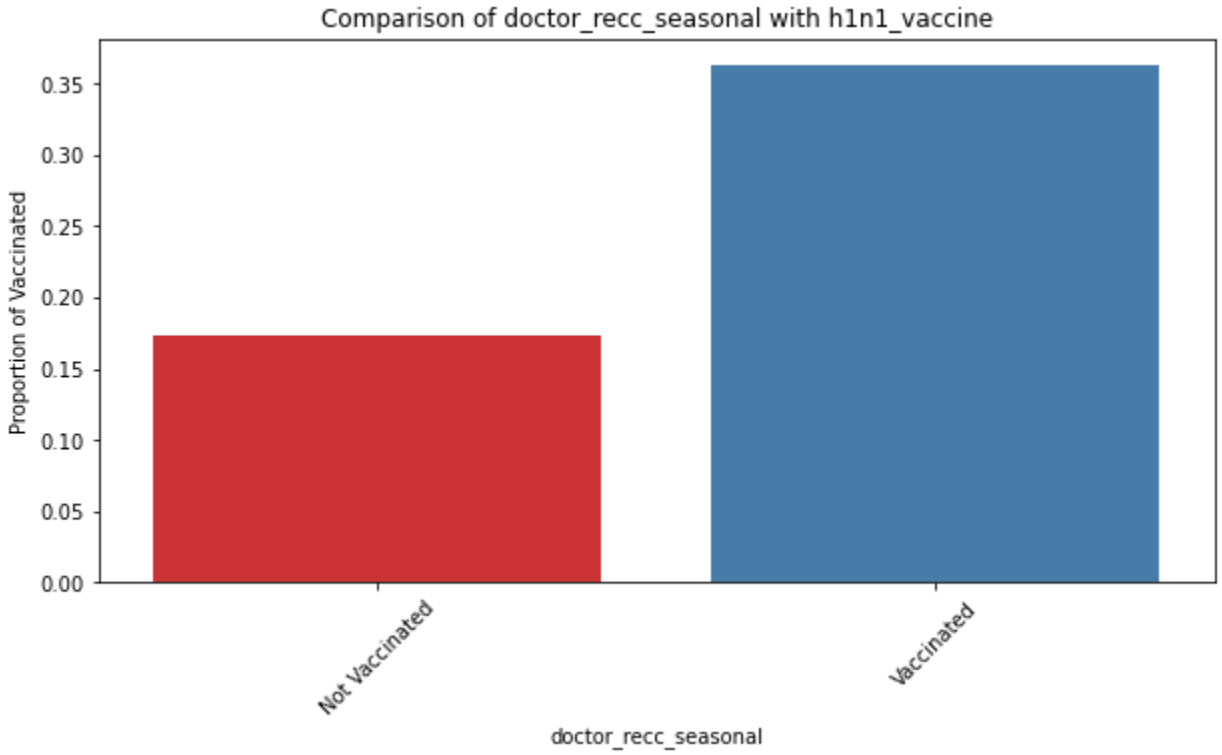

Figure 31: Doctor recommendation to get seasonal flu vaccine vs H1N1 vaccination

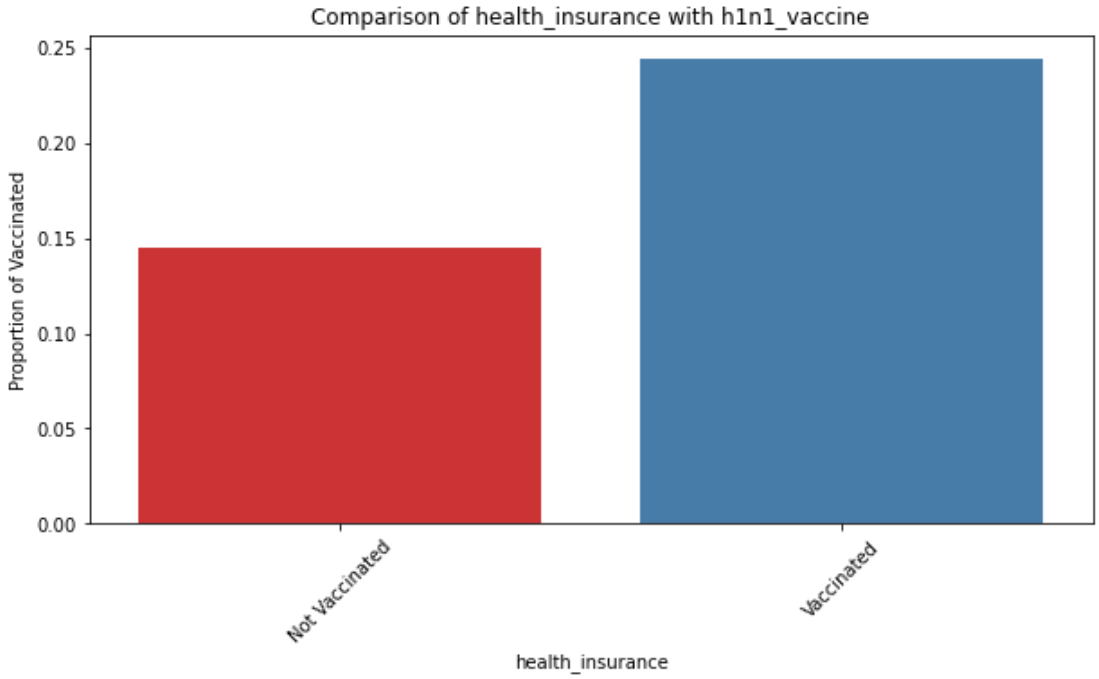

Figure 32: Health Insurance vs H1N1 vaccination

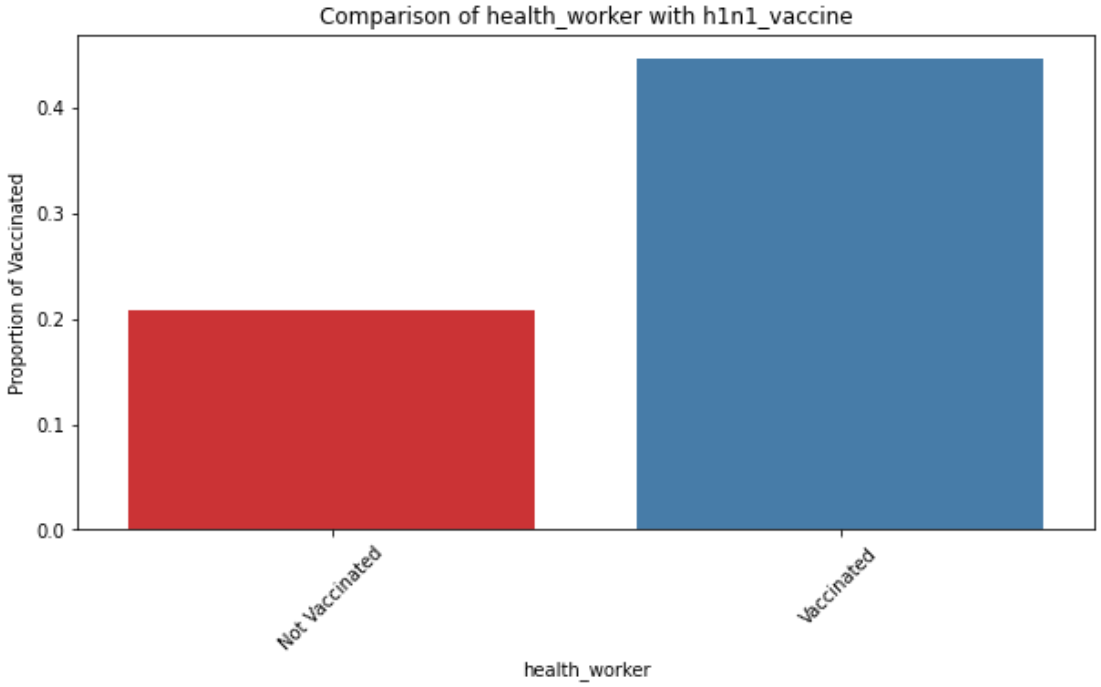

Figure 33: Health Worker vs H1N1 vaccination

Quite expected that **health related features** had played a major role in the number of those vaccinated.

Chronic illnesses, Doctors recommendations, health insurances and whether the respondent was a health worker all positively influenced vaccination.

Multivariate Analysis

Exploring the relationships between multiple features to identify patterns, correlations and interactions among them yielded the following results.

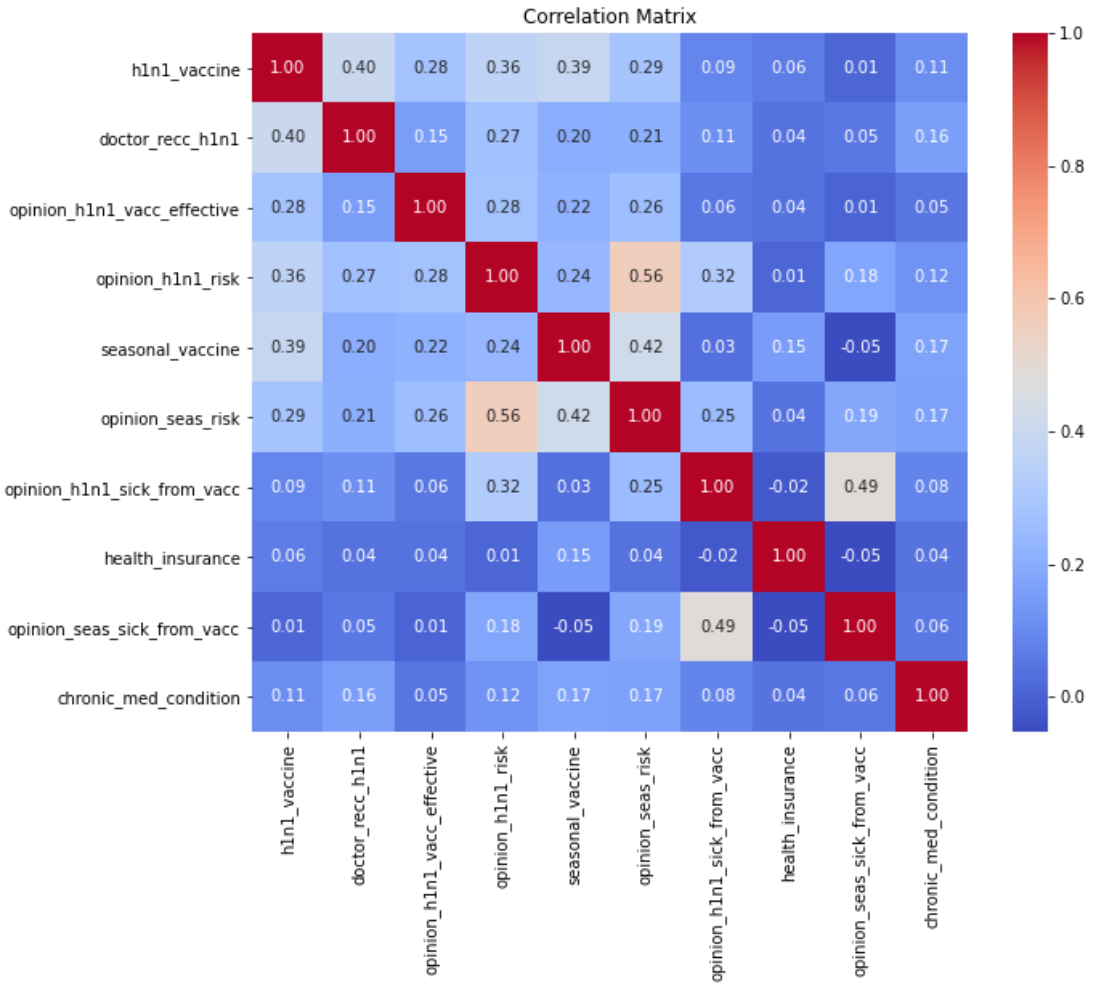

Figure 34: Correlation Matrix of multiple features

Features that seem to have been more highly correlated with **H1N1 vaccine** were, **doctor's recommendation, respondent's opinion on the effectiveness of the swine flu vaccine, respondent's opinion on whether the swine flu vaccine was risky, whether they received the seasonal flu vaccine as well and their opinion on the risks associated with the seasonal flu vaccine**.

MODELLING

This section dealt with building and training predictive models to determine the likelihood of H1N1 vaccine uptake based on various health, behavioral, demographic, and opinion features. This will help identify key factors influencing vaccine adoption and guide targeted public health interventions.

Two different types of classifiers were explored for this project. The first type was a **Logistic Regression** machine learning model from which a baseline model was trained and then a 2nd model was hyperparameter tuned, meaning it was tuned to improve its performance.

The second type of classifier was a **Decision Tree classifier** from which also 2 iterations were trained, the second one hyperparameter tuned from the 1st.

The first step was creating a new data frame that contained only the features that had proved to be most useful and highly correlated with the target feature (h1n1_vaccine) from the EDA. These features were: **'doctor_recc_h1n1', 'opinion_h1n1_vacc_effective', 'seasonal_vaccine', 'opinion_seas_vacc_effective','opinion_seas_risk','chronic_med_condition' and 'opinion_h1n1_risk.**

This new data frame contained the independent features and would be used to train the models to predict the dependent target feature.

Justification

Logistic Regression is a simple yet powerful baseline model that provides easily interpretable results.

Decision Trees are interpretable, non-linear models capable of capturing complex interactions between features. They would be useful for identifying subgroups that exhibit distinct vaccination behaviors.

Comparing the performance of linear and non-linear models helped identify the best approach for public health decision-making.

Logistic Regression Baseline Model

The performance metrics of the baseline model were as follows on the data it was tested on:

Precision: The model achieves a precision of ~71%, indicating that when the model predicts vaccine uptake, it is correct about 71% of the time.

Recall: The recall of ~54% shows that the model identifies just over half of the actual cases of vaccine uptake.

Accuracy: The overall accuracy of ~84% suggests good performance but might be inflated due to class imbalance.

F1-Score: The F1-score (~61%) balances precision and recall, providing a holistic measure of the model's predictive ability.

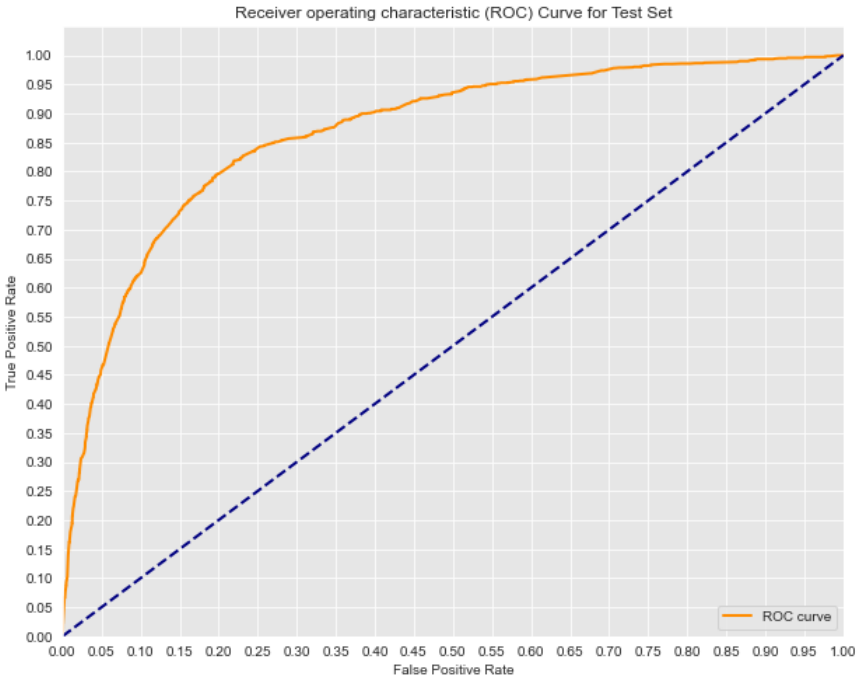

Figure 35:Logistic Regression baseline model AUC

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC), is the model's ability to distinguish between positive and negative classes. AUC of 0.867, indicates excellent discrimination. This suggests that the model performs well in distinguishing between the two classes (vaccine uptake vs. no vaccine uptake).

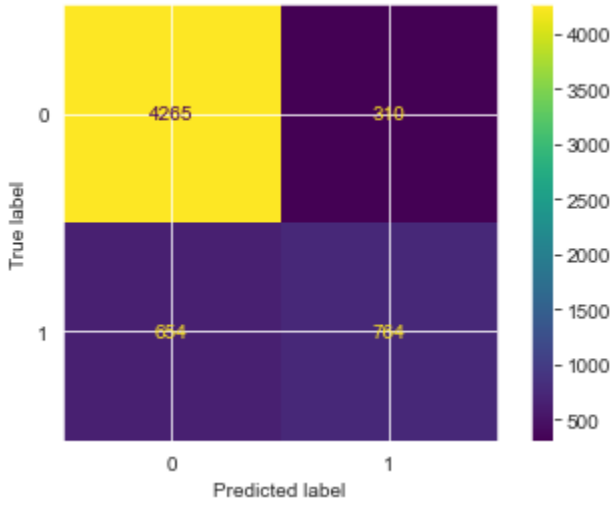

Figure 36: Logistic regression baseline model confusion matrix

True Negative Rate (Specificity): The model correctly identifies a high proportion of those who did not take the vaccine.

True Positive Rate (Sensitivity/Recall): The recall score aligns with the confusion matrix, showing the model captures 54% of actual vaccine uptake cases.

Strengths:

High specificity ensures that non-vaccine uptake is accurately predicted. The high AUC score suggests the model is strong at ranking predictions.

Weaknesses:

Moderate recall indicates that some vaccine uptake cases are missed.

Logistic Regression Tuned Model

The second model was tuned to improve recall and it showed significant changes in performance compared to the baseline model. Its performance metrics were as follows on the data it was tested on:

Precision: Decreased from 71% to 55%.

The model is less precise, meaning a higher proportion of predicted positive cases are incorrect. However, this is acceptable since the primary goal is to improve recall.

Recall: Increased from 54% to 80%.

The model now identifies a much larger proportion of actual positive cases, aligning with the project's focus on recall to minimize missed high-risk individuals.

Accuracy: Slightly decreased from 84% to 80%.

Accuracy dropped slightly, likely due to the trade-off between precision and recall.

F1 Score: Increased from 61% to 65%.

The F1 score improved, reflecting a better balance between precision and recall.

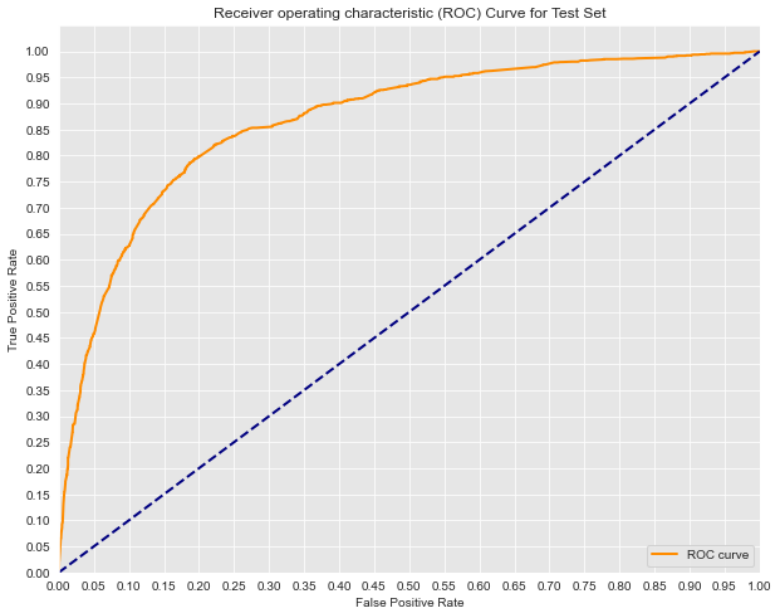

Figure 37: Tuned Regression baseline model AUC

The new AUC of 0.8669 suggests that the adjustments made of stronger regularization and balanced classes did not negatively impact the model's overall performance in terms of its ability to distinguish between classes.

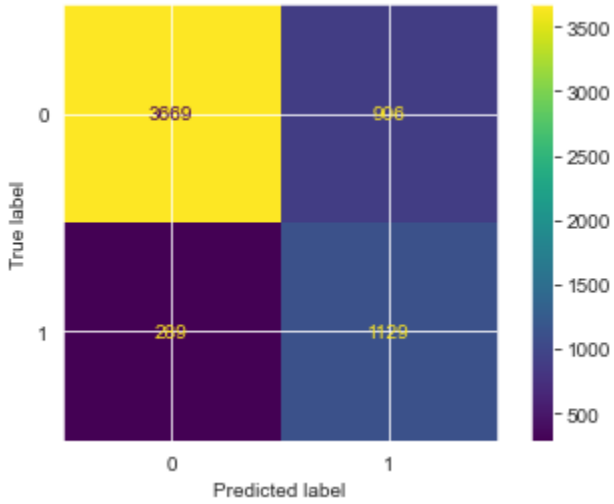

Figure 38: Tuned Regression baseline model confusion matrix

The model improved its recall (detecting more true positives) compared to the baseline model, indicating better performance in capturing individuals who actually took the vaccine.

Decision Tree Classifier

After training a decision tree classifier on the same data as the logistic regression models, these were the performance metrics of the 1st decision tree classifier:

Training data precision = 76%

Test data precision = 67%

Training data recall = 60%

Test data recall = 53%

Training data accuracy = 86%

Test data accuracy = 83%

Training data F1 score = 66%

Test data F1 score = 59%

The model performed better on the training data that in did on the test data in all metrics meaning it might be suffering from overfitting.

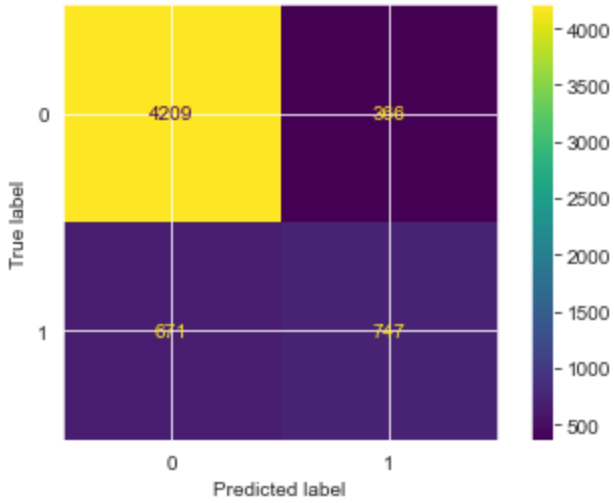

Figure 39: Decision Tree model confusion matrix

It had an AUC score of 0.82 indicating a good level of discrimination.

Tuned Decision Tree Classifier

To attempt and fix the overfitting issues with the first decision tree model, some hyperparameter tuning was done and these were the results after tuning.

Recall: Significant Increase: The recall for the tuned model (78% on the test set) is much higher compared to the original model (0.5282).

Precision: Slight Decrease: The test precision for the tuned model (53%) is lower than the original (67%). This trade-off is expected when prioritizing recall.

F1 Score: Improved: The F1 score of the tuned model (63%) is higher than that of the original model (59%). A better balance between precision and recall.

Accuracy: Slight Drop: The accuracy of the tuned model (78%) is lower than the original model (83%)

AUC: Improved: The AUC for the tuned model (84%) is slightly higher than the original (82%), indicating that the overall discriminatory power.

EVALUATION

For a project aimed at **guiding public health efforts and increasing vaccine uptake**, the most important performance metrics should focus on the model's ability to identify individuals likely to vaccinate or resist vaccination, while minimizing misclassification that could misguide outreach strategies.

Hence, **Recall** the ability of the model to correctly identify individuals who are likely to vaccinate, **AUC** the model's ability to distinguish between vaccinated and non-vaccinated individuals across different thresholds and **Precision** which measures the proportion of predicted vaccinated individuals who are actually vaccinated, are the most important metrics of success.

Tuned Logistic Regression Model vs Tuned Decision Trees Classifier

Recall:

The tuned logistic regression model has a slightly higher recall (80%) compared to the tuned decision tree model (76%). This indicates that it is better at identifying all actual positive cases.

AUC:

The tuned logistic regression model has an AUC of 87%, which is higher than the decision tree's AUC of 84%. This suggests that it is better at distinguishing between positive and negative classes, making it more effective for decision-making.

Precision:

The tuned logistic regression model has a higher test precision (56%) than the tuned decision tree model (53%). This means it is more accurate when predicting positive cases.

CONCLUSIONS

Exploratory Data Analysis:

EDA revealed key trends and relationships between the features and the target variable, which allowed for targeted feature engineering and preprocessing.

Variables such as **doctor_recc_h1n1, opinion_h1n1_vacc_effective, seasonal_vaccine** and **chronic_med_condition** were identified as significant predictors for the likelihood of individuals receiving the H1N1 vaccine.

Modelling Results:

Multiple models were evaluated to find the best fit for predicting the likelihood of H1N1 vaccine uptake.

The tuned logistic regression model emerged as the most effective model, making it suitable for real-world deployment where understanding both positive and negative cases is crucial. The decision tree model, despite being optimized, had lower performance metrics.

RECOMMENDATIONS

Model Selection:

Choose the tuned logistic regression model as the final model for implementation due to its superior performance metrics. It ensures robust prediction performance and is highly interpretable, which is vital for public health applications where stakeholders need to understand model outcomes.

Public Health Strategy:

Use the insights from the model to tailor public health campaigns and interventions, focusing efforts on communities and demographics identified as being at higher risk of not receiving the vaccine.

Employ targeted messaging that aligns with the opinions and perceptions about vaccine effectiveness.

Enhance Public Health Campaigns focused on Doctor Recommendations. The EDA highlighted variables such as **doctor_recc_h1n1** were significant predictors of vaccine uptake. Public health officials should leverage this insight to launch campaigns that encourage healthcare providers to actively recommend the H1N1 vaccine to patients.

NEXT STEPS

Analyze the factors that led to high adoption rates of Seasonal Flu vaccination where EDA showed that an almost equal number of respondents received the vaccination as to those who did not receive it.

Collect more data or improve data quality to enhance model performance, particularly for features that showed lower impact during analysis.

Monitor model performance over time to adapt to changes in public behavior or new health information, ensuring long lasting effectiveness.