Factors Influencing H1N1 Vaccine Uptake: A Predictive Analysis

Business Understanding

Vaccines stimulate the immune system to protect against specific diseases by containing parts of pathogens. Flu vaccines, including seasonal and H1N1 vaccines, help prevent the spread of influenza. Seasonal vaccines are updated yearly, while the H1N1 vaccine was developed after the 2009 pandemic. Vaccination protects individuals and promotes herd immunity, but challenges such as vaccine hesitancy, misinformation, and limited access can hinder widespread uptake. This project aims to predict factors influencing H1N1 vaccine uptake using data from the National 2009 H1N1 Flu Survey. The primary stakeholders are public health officials and policymakers, who will use the insights to design targeted vaccination strategies and improve future public health campaigns. The project will focus on predicting vaccine uptake based on demographic, behavioral, and health-related factors, and will not address issues related to vaccine distribution, policy, or effectiveness. The data source includes demographic information, health behaviors, and vaccination history. The project will be completed within a week, with the goal of providing actionable insights for improving public health outcomes. Clear alignment with stakeholders is essential to ensure the project meets their expectations.

Data Understanding

The data for this project comes from three datasets: training_set_features.csv, test_set_features.csv, and training_set_labels.csv. The target variables are whether individuals received the H1N1 vaccine or the seasonal flu vaccine, while the predictors include demographic, health-related, and behavioral features such as age, income, health concerns, and vaccination recommendations. The data includes both categorical (e.g., age group, marital status) and numerical (e.g., household size, income) variables, with some binary and ordinal features. The dataset's size and distribution will be examined during exploration, and if necessary, resampling techniques may be applied to address imbalances. The data is collected via surveys, but it may contain biases or missing values, which will require cleaning and preprocessing before building the model.

Data Preparation

Data preprocessing included:

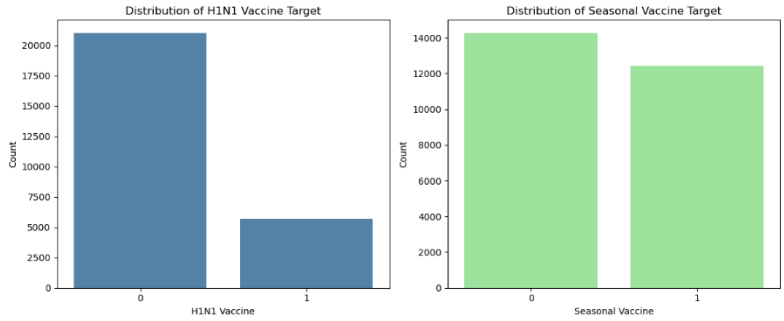
- **Handling Missing Values**: Missing values were imputed or removed based on the type of variable and the number of missing entries.
- **Encoding Categorical Variables**: Categorical features such as gender, vaccine perception, and health conditions were one-hot encoded to convert them into numerical format.
- **Checking for Duplicates**: By identifying and removing duplicates, we ensure the dataset's integrity and prevent redundancy from influencing the results

Data Visualizations

Several key visualizations were created to better understand the data and help in the model-building process:

Distribution of Target Variables

Purpose: Understand the balance of target variables (h1n1_vaccine and seasonal vaccine).



H1N1 Vaccine Distribution:

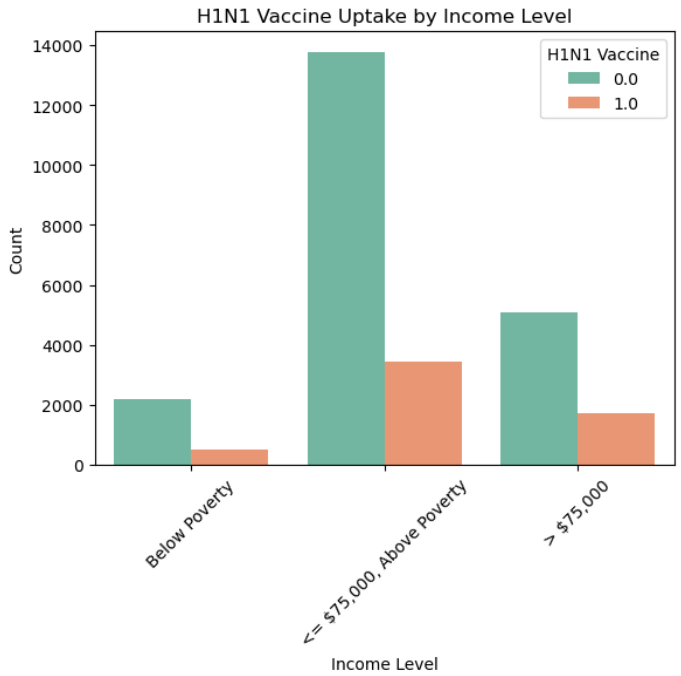
The majority of respondents (label 0) did not receive the H1N1 vaccine. A smaller proportion of respondents (label 1) received the vaccine. This indicates an imbalance in the target variable, as the number of people who did not take the H1N1 vaccine is significantly higher than those who did.

Seasonal Vaccine Distribution:

The distribution between those who did not receive the seasonal vaccine (label 0) and those who did (label 1) is more balanced compared to the H1N1 vaccine distribution. There is still a slightly higher number of respondents who did not receive the seasonal vaccine, but the difference is less pronounced

Vaccine Uptake by Income Level

Purpose: Understand how income poverty influences vaccine uptake.



<= $75,000, Above Poverty: This income category has the highest overall count of individuals. A significant majority in this group did not take the H1N1 vaccine (0.0). The number of individuals who took the vaccine (1.0) is comparatively lower.

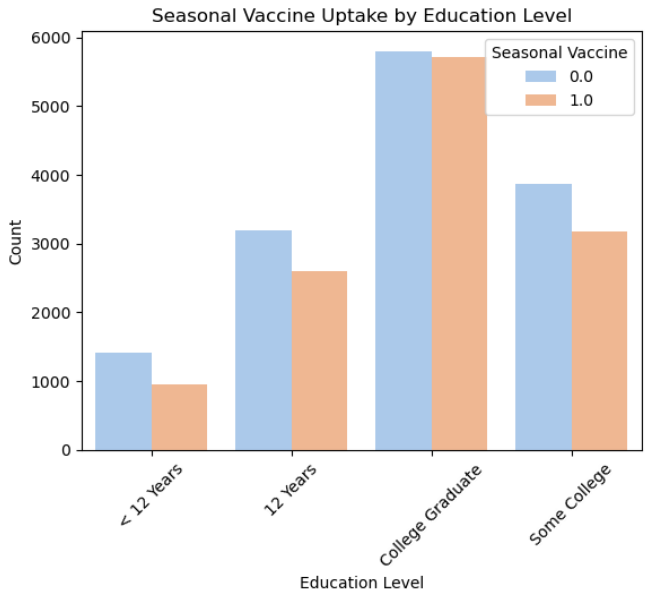
Below poverty: This group has the lowest total count of individuals. The uptake of the vaccine (1.0) is low but more proportionate to those who did not take it (0.0) compared to other income groups.

> $75,000:This group has a smaller total count than the middle-income category. Similar to the other groups, a larger proportion of individuals did not take the vaccine (0.0), though the uptake (1.0) is slightly higher than in the lowest income group

The middle-income group (<= $75,000, Above Poverty) is the largest and shows the lowest vaccine uptake proportionally. This could indicate a need for targeted interventions in this demographic. Both the lowest and highest income groups show relatively lower H1N1 vaccine uptake, but the absolute count in the lowest income group is much smaller, suggesting additional barriers.

Vaccine Uptake by Education Level

Purpose: Assess the impact of education on vaccine uptake



College Graduates: This group has the highest count of individuals overall. A significant proportion of college graduates took the seasonal vaccine (1.0).

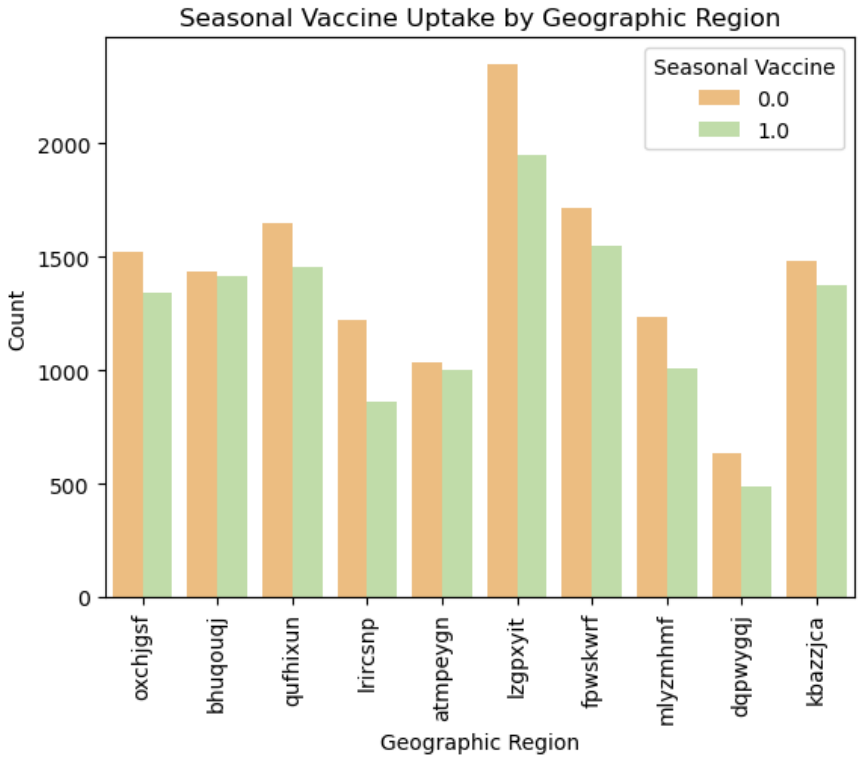
Education Levels Below College: As education levels decrease (e.g., <12 Years, 12 Years), the total count of individuals in these groups declines. The proportion of individuals taking the vaccine (1.0) is lower compared to those who did not (0.0).

Some College: This group has a balanced representation but still shows a higher count of individuals not taking the vaccine.

Higher education levels, particularly college graduates, are positively associated with seasonal vaccine uptake. Individuals with lower education levels may have barriers to vaccine uptake, such as access, awareness, or beliefs. This could be an area for targeted public health campaigns.

Geographic Region Analysis

Purpose: Visualize how vaccine uptake varies across hhs_geo_region.



In most regions, the count of individuals who did not take the vaccine (orange bars) is higher than those who did (green bars).

The region "lzpxvyit" stands out with the highest number of individuals who did not take the vaccine (0.0) and also the highest uptake of the vaccine (1.0), indicating a large population in this region.

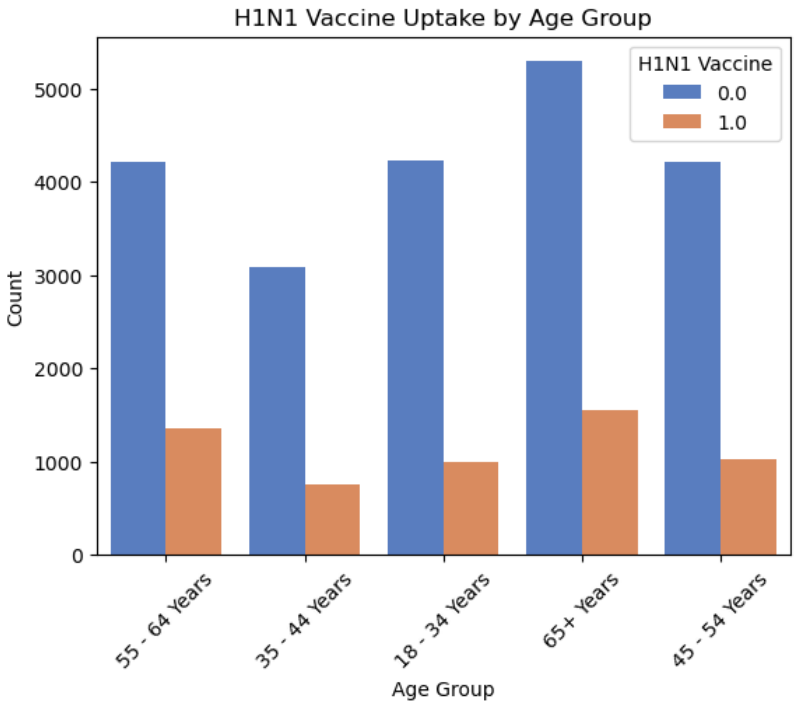
Regions like "fpwskwrf" and "mllyzmhmf" show lower vaccine uptake compared to other regions.

The counts for non-uptake (0.0) and uptake (1.0) appear to be more balanced in some regions, such as "bhugqouj" and "kbazzjca."

Overall, the chart suggests significant variation in seasonal flu vaccine uptake across geographic regions, with non-uptake generally outpacing uptake.

Age Group vs. Vaccination

Purpose: Analyze how vaccination uptake differs by age group.



For all age groups, the count of individuals who did not take the vaccine (blue bars) is significantly higher than those who did (orange bars).

The 65+ years age group has the highest number of individuals represented, both for vaccine uptake (1.0) and non-uptake (0.0).

The youngest age group (18–34 years) shows the lowest uptake of the vaccine (orange bar) relative to the others.

This chart suggests that vaccine uptake was generally low across all age groups, with older populations (especially 65+ years) having the highest representation overall.

Feature Encoding

Convert categorical variables into a numeric format to make the data usable by machine learning algorithms. By applying techniques such as one-hot encoding or label encoding, we ensure that categorical variables are properly represented, enabling the model to process and learn from them effectively.

Modeling

Two models were implemented to predict vaccine uptake: Logistic Regression and Decision Trees. Each model was evaluated using accuracy, precision, recall, F1-score, and a confusion matrix. I also performed

hyperparameter tuning for both the models to identify the best settings for parameters like maximum depth and minimum sample split.

- **Logistic Regression**: A basic linear model that predicted vaccine uptake based on the input features. This model provided a good baseline but struggled with the class imbalance. Hyperparameter tuning improved its performance, resulting in a deeper tree with better classification ability
- **Decision Tree**: A non-linear model that split the data at various decision points based on feature values. Hyperparameter tuning improved its performance, resulting in a deeper tree with better classification ability.

Evaluation

Model performance was evaluated using a classification report, which provided the following results .

Simple Logistic Regression Model:

- **Accuracy**: 79%
- **Precision (Class 0)**: 0.86
- **Recall (Class 0)**: 0.87
- **Precision (Class 1)**: 0.50
- **Recall (Class 1)**: 0.48
- **F1-Score (Class 1)**: 0.49

The logistic regression model performed reasonably well in predicting non-vaccinated individuals (class 0) but struggled with vaccinated individuals (class 1), resulting in a low recall for class 1.

Simple Decision Tree Model:

- **Accuracy**: 76%
- **Precision (Class 0)**: 0.83
- **Recall (Class 0)**: 0.89
- **Precision (Class 1)**: 0.42
- **Recall (Class 1)**: 0.30
- **F1-Score (Class 1)**: 0.35

The simple decision tree model performed similarly to logistic regression in terms of predicting non-vaccinated individuals but was even worse at predicting vaccinated individuals. Its low recall for class 1 indicates it misses a significant portion of vaccinated people.

Tuned Logistic Regression Model:

- **Accuracy**: 81%
- **Precision (Class 0)**: 0.89
- **Recall (Class 0)**: 0.94
- **Precision (Class 1)**: 0.59
- **Recall (Class 1)**: 0.33
- **F1-Score (Class 1)**: 0.42

With hyperparameter tuning, logistic regression saw a boost in accuracy, especially in predicting non-vaccinated individuals. However, the recall for class 1 still remained low, and the model still struggled with predicting vaccinated individuals.

Tuned Decision Tree Model:

- **Accuracy**: 81%
- **Precision (Class 0)**: 0.84
- **Recall (Class 0)**: 0.94
- **Precision (Class 1)**: 0.59
- **Recall (Class 1)**: 0.33
- **F1-Score (Class 1)**: 0.42

After tuning, the decision tree model showed similar performance to the tuned logistic regression model. Both models had high precision and recall for non-vaccinated individuals but struggled with the prediction of vaccinated individuals

From the confusion matrix for all four models, we observed that:

- **Non-vaccinated individuals (class 0)** were correctly identified with high precision and recall in all models.
- **Vaccinated individuals (class 1)** were harder to predict, with models generating a higher number of **false negatives**, meaning they missed many vaccinated individuals.

The model predicts the likelihood of vaccine uptake based on various factors such as age, gender, health status, and prior vaccine perceptions. The most important features influencing vaccine uptake include health-related factors (such as chronic conditions) and personal beliefs about vaccine effectiveness. While the model is able to predict the target with reasonable accuracy, it is more effective for certain groups (e.g., those with strong health concerns) and less reliable for others (e.g., those with vaccine hesitancy). To improve predictions and outcomes, the business could focus on addressing concerns related to vaccine effectiveness and targeting high-risk groups more effectively.

Recommendations

The model's predictions are useful for targeting specific groups to increase vaccine uptake, efficiently allocating resources, and designing personalized campaigns. However, it may not be useful in situations where the data is incomplete or unrepresentative of certain populations, or if external factors change drastically. To improve results, the business could enhance data quality, focus on specific demographics, and update the model regularly with new data. By adjusting input variables, such as adding features related to health behavior or socioeconomic status, the model can better predict vaccine uptake and guide more effective interventions.

The models built, including Logistic Regression and Decision Tree, have shown decent performance with an accuracy of around 81%. While there is room for improvement, particularly in predicting certain groups who are less likely to take the vaccine, the current results provide valuable insights. To further refine the models, we could enhance feature engineering, adjust hyperparameters, or explore advanced methods like ensemble models. However, the model's performance is sufficient to move forward with deployment. The insights gathered so far can be used to guide strategies aimed at improving vaccine uptake, making it ready for practical use.