Exploring Determinants of H1N1 Vaccine Uptake: A Data-Driven Public Health Approach Using Classification Models

Background Information

1. What happened in 2009 that shook the world?

In the early months of the 2009, the first case of H1N1 virus (also called Swine flu) was recorded in Mexico and then due to the airline travel, the virus later rapidly all spread around the world and later was declared a pandemic by the World Health Organization. The virus caused an estimated 284,400 deaths globally and was found that this 2009 H1N1 strain is a unique combination of human, swine and avian(bird) influenza A viruses.

Children and young adults became the most infected and affected. However, it was found that complications leading to hospitalization and the need for intensive care were prevalent in :

- Very young children
- Pregnant women
- Those who are morbidly obesity
- Those with underlying medical conditions such as chronic lung and cardiac diseases, diabetes, weak immunity
- Those with Bacterial coinfection were among the fatal cases

The 2009 H1N1 influenza pandemic highlighted the importance of understanding vaccination patterns to improve public health.

2. What is H1N1 Flu?

The H1N1 flu, also called the swine flu, is a type of influenza A virus.

Symptoms usually start quickly and can include:

- Fever for some patients
- Muscle Aches
- Chills and sweating
- Cough
- Sore throat
- Runny or stuffy nose
- Watery and even red eyes
- Eye pain

- Body aches
- Headache
- Tiredness and general weakness
- Diarrhea
- Stomach problems, vomiting which is more common in children

These swine flu symptoms develop about 1 to 4 days after exposure to the virus.

With time, the H1N1 flu strain from the pandemic became one of the strains that cause seasonal flu. Most people with the flu get better on their own but the complications are much worse and can be much more deadly, especially for people at high risk.

To curb this type of flu, the seasonal flu vaccine is used to protect against the H1N1 flu and other seasonal flu viruses.

Problem Statement

Immunization/Vaccination is thus a matter that cannot be ignored as an important tool in managing the spread of influenza.

As seen during the COVID-19 pandemic, it was clear that personal vaccination decisions are influenced by multiple factors, including background, health behaviour and opinions toward vaccines.

The National 2009 H1N1 Flu Survey provides rich data for analyzing these factors and how they influence vaccination uptake.

This analysis should support public health experts on what to do to provide better vaccination outreach.

Objectives

- The **main objective** is building a predictive model that can accurately forecast whether a person received the H1N1 vaccine based on a variety of features such as their concerns about H1N1, health behaviors, opinions about the vaccine, demographics, and more.

- Analyse the various factors through visualizations to analyse the behaviour of each when it comes to receiving vaccinations. For example, to see how various opinions are distributed.

- Evaluate the predictive model and provide actionable insights to inform future public health vaccination strategies.

Metrics of Success

Accuracy: Percentage of correct predictions for vaccination status.
Precision: Ability to correctly identify vaccinated individuals.
Recall (Sensitivity): Ability to detect all vaccinated individuals.
F1 Score: Balance between precision and recall, especially if the dataset is imbalanced.
Area Under the ROC Curve (AUC-ROC): Overall ability of the model to distinguish between vaccinated and unvaccinated individuals.

Data Understanding

Data source and guidelines of the use of this data

The data for this competition comes from the National 2009 H1N1 Flu Survey (NHFS).

In their own words:
The National 2009 H1N1 Flu Survey (NHFS) was sponsored by the National Center for Immunization and Respiratory Diseases (NCIRD) and conducted jointly by NCIRD and the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The NHFS was a list-assisted random-digit-dialing telephone survey of households, designed to monitor influenza immunization coverage in the 2009-10 season.

The target population for the NHFS was all persons 6 months or older living in the United States at the time of the interview. Data from the NHFS were used to produce timely estimates of vaccination coverage rates for both the monovalent pH1N1 and trivalent seasonal influenza vaccines.

The NHFS was conducted between October 2009 and June 2010. It was one-time survey designed specifically to monitor vaccination during the 2009-2010 flu season in response to the 2009 H1N1 pandemic. The CDC has other ongoing programs for annual phone surveys that continue to monitor seasonal flu vaccination.

The source dataset comes with the following data use restrictions:

- The Public Health Service Act (Section 308(d)) provides that the data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and

Prevention (CDC), may be used only for the purpose of health statistical reporting and analysis.

- Any effort to determine the identity of any reported case is prohibited by this law.

- NCHS does all it can to ensure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted from the data files. Any intentional identification or disclosure of a person or establishment violates the assurances of confidentiality given to the providers of the information.

Therefore, users will:

- Use the data in these data files for statistical reporting and analysis only. Make no use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS, of any such discovery (1 (800) 232-4636). Not link these data files with individually identifiable data from other NCHS or non-NCHS data files. By using these data, you signify your agreement to comply with the above requirements.

Link:

https://webarchive.loc.gov/all/20140511031000/http://www.cdc.gov/nchs/nis/about_nis.htm#h1n1

The data to be used for this analysis is broken down into four parts. Below is a breakdown of what each type of data provides:

1. Training set features - Dataset contains the features associated with each respondent in the training data
2. Training set labels - A unique identifier for each respondent in the training set.
3. Test set features - Contains features for respondents in the test set. This is what you'll use to make predictions.
4. Submission Format - This is the format in which you need to submit your predictions

Columns and what the meaning

For the Submission Format data and the Training set labels, the have the same columns as follows:

The other remaining data sets also have the same columns (Training set features and Test set features data sets). These 35 columns are as follows:
Note that for all binary variables: 0 = No; 1 = Yes.

Here's the numbered list of feature descriptions for easy copying into your Word document:

Feature Descriptions

1. **respondent_id**
   A unique and random identifier.
2. **h1n1_concern**
   Level of concern about the H1N1 flu.
   - 0 = Not at all concerned
   - 1 = Not very concerned
   - 2 = Somewhat concerned
   - 3 = Very concerned
3. **h1n1_knowledge**
   Level of knowledge about H1N1 flu.
   - 0 = No knowledge
   - 1 = A little knowledge
   - 2 = A lot of knowledge
4. **behavioral_antiviral_meds**
   Has taken antiviral medications. (binary)
5. **behavioral_avoidance**
   Has avoided close contact with others with flu-like symptoms. (binary)
6. **behavioral_face_mask**
   Has bought a face mask. (binary)
7. **behavioral_wash_hands**
   Has frequently washed hands or used hand sanitizer. (binary)
8. **behavioral_large_gatherings**
   Has reduced time at large gatherings. (binary)
9. **behavioral_outside_home**
   Has reduced contact with people outside of own household. (binary)
10. **behavioral_touch_face**
    Has avoided touching eyes, nose, or mouth. (binary)
11. **doctor_recc_h1n1**
    H1N1 flu vaccine was recommended by doctor. (binary)
12. **doctor_recc_seasonal**
    Seasonal flu vaccine was recommended by doctor. (binary)
13. **chronic_med_condition**
    Has any of the following chronic medical conditions: asthma or other lung conditions, diabetes, heart conditions, kidney conditions, sickle cell anemia, neurological or

neuromuscular conditions, liver conditions, or weakened immune systems due to chronic illnesses or medications. (binary)

14. **child_under_6_months**
    Has regular close contact with a child under the age of six months. (binary)
15. **health_worker**
    Is a healthcare worker. (binary)
16. **health_insurance**
    Has health insurance. (binary)
17. **opinion_h1n1_vacc_effective**
    Respondent's opinion about H1N1 vaccine effectiveness.
    - 1 = Not at all effective
    - 2 = Not very effective
    - 3 = Don’t know
    - 4 = Somewhat effective
    - 5 = Very effective
18. **opinion_h1n1_risk**
    Respondent's opinion about the risk of getting sick with H1N1 flu without a vaccine.
    - 1 = Very Low
    - 2 = Somewhat low
    - 3 = Don’t know
    - 4 = Somewhat high
    - 5 = Very high
19. **opinion_h1n1_sick_from_vacc**
    Respondent's worry of getting sick from taking the H1N1 vaccine.
    - 1 = Not at all worried
    - 2 = Not very worried
    - 3 = Don’t know
    - 4 = Somewhat worried
    - 5 = Very worried
20. **opinion_seas_vacc_effective**
    Respondent's opinion about seasonal flu vaccine effectiveness.
    - 1 = Not at all effective
    - 2 = Not very effective
    - 3 = Don’t know
    - 4 = Somewhat effective
    - 5 = Very effective
21. **opinion_seas_risk**
    Respondent's opinion about the risk of getting sick with seasonal flu without a vaccine.
    - 1 = Very Low
    - 2 = Somewhat low
    - 3 = Don’t know
    - 4 = Somewhat high
    - 5 = Very high
22. **opinion_seas_sick_from_vacc**
    Respondent's worry of getting sick from taking the seasonal flu vaccine.
    - 1 = Not at all worried

    - 2 = Not very worried
    - 3 = Don’t know
    - 4 = Somewhat worried
    - 5 = Very worried
23. **age_group**
    Age group of the respondent.
24. **education**
    Self-reported education level.
25. **race**
    Race of the respondent.
26. **sex**
    Sex of the respondent.
27. **income_poverty**
    Household annual income of the respondent with respect to 2008 Census poverty thresholds.
28. **marital_status**
    Marital status of the respondent.
29. **rent_or_own**
    Housing situation of the respondent.
30. **employment_status**
    Employment status of the respondent.
31. **hhs_geo_region**
    Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
32. **census_msa**
    Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
33. **household_adults**
    Number of other adults in the household, top-coded to 3.
34. **household_children**
    Number of children in the household, top-coded to 3.
35. **employment_industry**
    Type of industry the respondent is employed in. Values are represented as short random character strings.
36. **employment_occupation**
    Type of occupation of the respondent. Values are represented as short random character strings.

The data above can be grouped into:

- **Demographics**
  These columns describe basic demographic information about the respondents.
  - age_group, education, race, sex, income_poverty, marital_status, etc.
- **Health Behaviors**
  These columns represent various health-related behaviors and actions taken by the respondents to prevent illness.
  - behavioral_avoidance, behavioral_face_mask, behavioral_touch_face, etc.
- **Health Status**
  These columns provide information about the health conditions and situations of the respondents.
  - chronic_med_condition, child_under_6_months, health_worker, health_insurance.
- **Opinions About Vaccines**
  These columns reflect the respondents' views on the effectiveness and risks associated with the H1N1 and seasonal flu vaccines.
  - opinion_h1n1_vacc_effective, opinion_h1n1_risk, opinion_h1n1_sick_from_vacc, opinion_seas_vacc_effective, opinion_seas_risk, opinion_seas_sick_from_vacc.
- **Doctor Recommendations**
  These columns indicate whether the respondents' doctors recommended them to get the H1N1 or seasonal flu vaccine.
  - doctor_recc_h1n1, doctor_recc_seasonal.
- **Geographic and Employment Data**
  These columns provide information about where the respondents live and their employment details.
  - hhs_geo_region, employment_status, employment_industry, census_msa.
- **Respondent ID**
  This column is a unique identifier for each respondent, linking these features to other datasets.

Data Understanding

Handling of missing data

The only two data sets that had missing value were:

- Training set features
- Testing set features

The % of missing data is as shown below:

- Training set features

```
child_under_6_months            3.07%
health_worker                   3.01%
health_insurance               45.96%
opinion_h1n1_vacc_effective     1.46%
opinion_h1n1_risk               1.45%
opinion_h1n1_sick_from_vacc     1.48%
opinion_seas_vacc_effective     1.73%
opinion_seas_risk               1.92%
opinion_seas_sick_from_vacc     2.01%
age_group                       0.00%
education                       5.27%
race                            0.00%
sex                             0.00%
income_poverty                 16.56%
marital_status                  5.27%
rent_or_own                     7.65%
employment_status               5.48%
hhs_geo_region                  0.00%
census_msa                      0.00%
household_adults                0.93%
household_children              0.93%
employment_industry            49.91%
employment_occupation          50.44%
dtype: object
========================================
Number of Duplicate Rows: 0
```

- Testing set features

```
child_under_6_months            3.04%
health_worker                   2.95%
health_insurance               45.78%
opinion_h1n1_vacc_effective     1.49%
opinion_h1n1_risk               1.42%
opinion_h1n1_sick_from_vacc     1.40%
opinion_seas_vacc_effective     1.69%
opinion_seas_risk               1.87%
opinion_seas_sick_from_vacc     1.95%
age_group                       0.00%
education                       5.27%
race                            0.00%
sex                             0.00%
income_poverty                 16.84%
marital_status                  5.40%
rent_or_own                     7.62%
employment_status               5.51%
hhs_geo_region                  0.00%
census_msa                      0.00%
household_adults                0.84%
household_children              0.84%
employment_industry            49.70%
employment_occupation          50.27%
dtype: object
========================================
Number of Duplicate Rows: 0
```

Missing values were replaced by the modal category in each column where there were missing values.

Columns that had over 45% of missing values were dropped from the data.

Handling of duplicates

There were no duplicates in the data sets.

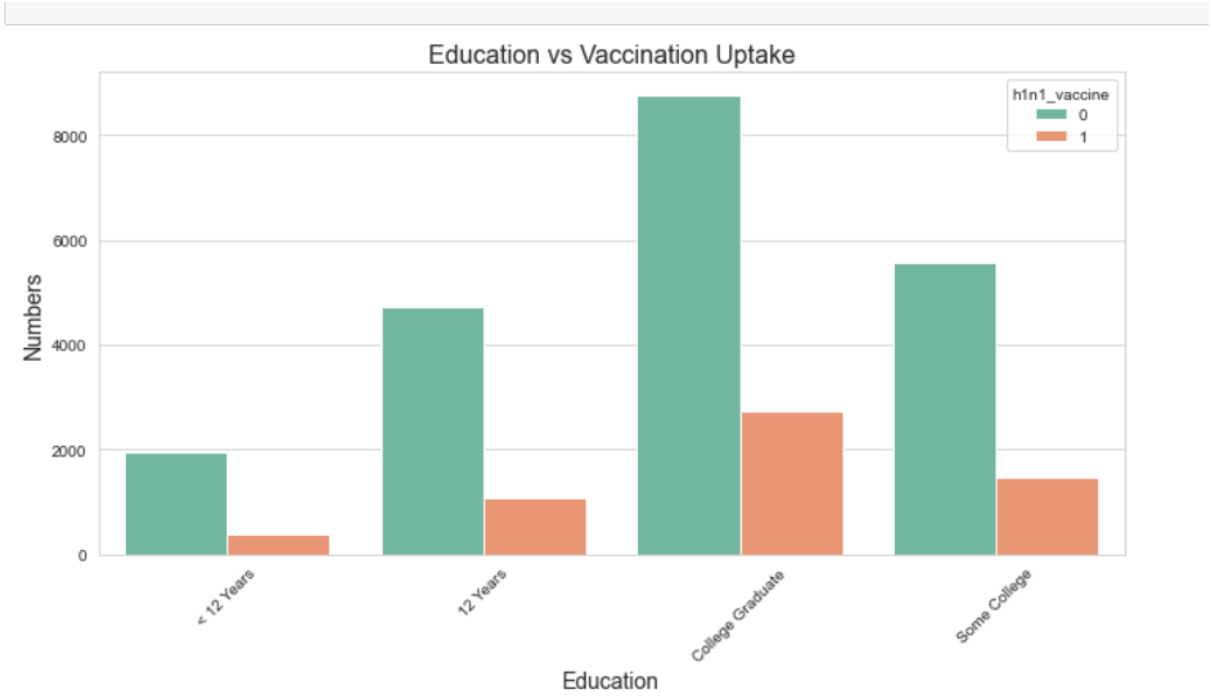
Handling of categorical data to be used for logistic regression

One hot encoding was applied to remove the categorical groupings to thus replace them with 0s and 1s

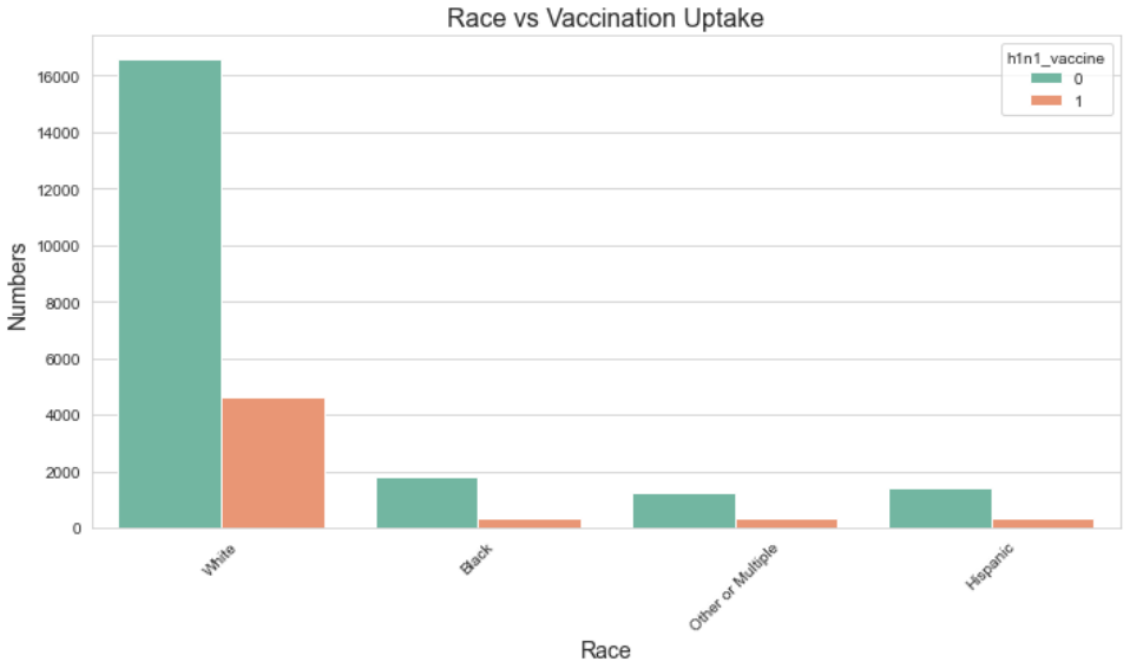
Data Analysis

In this part:

An analysis and visualization on some relevant demographics feature influence on vaccination uptake
For example: Education, Age, Sex

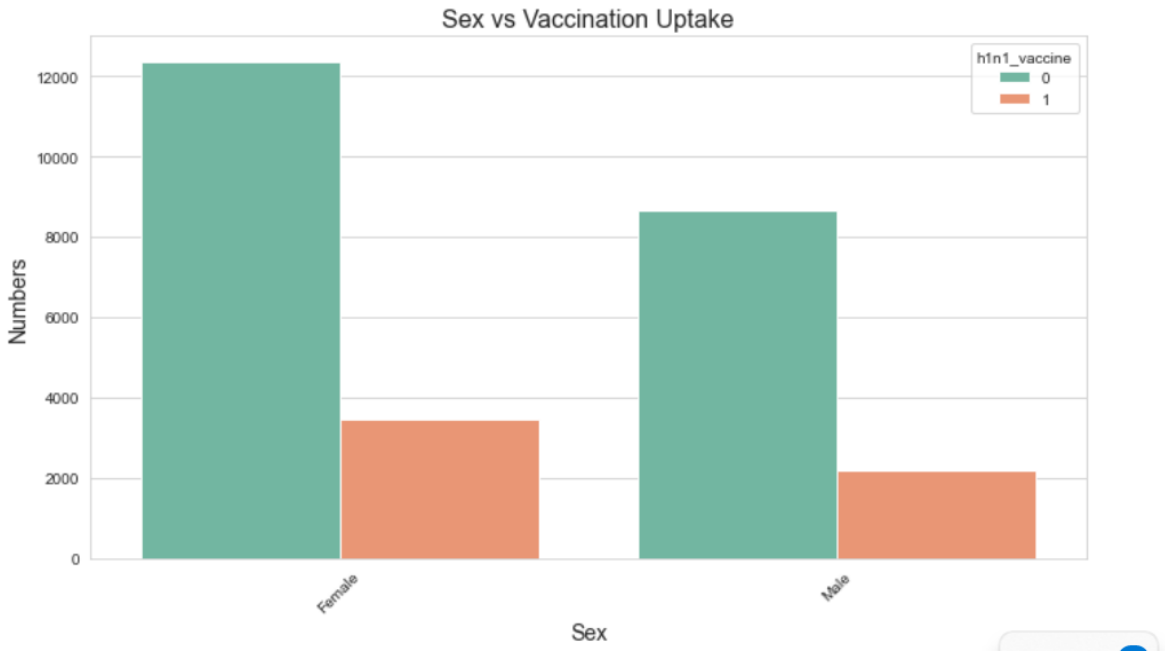


- From the graph we see when it came to education more college students did not take the vaccine



- From the graph we see that when it comes to race more white people are likely to get vaccinated and not get vaccinated. This also shows that the whites are the majority class and there the results represent this expectation.

- From the graph below see that when it comes to gender more women than men avoid getting vaccinated.



MODELING

Model Overview and Optimization

Logistic Regression was used as the baseline model for predicting the h1n1_vaccine target variable. Given the large number of features (around 65 after encoding and removing highly correlated ones), the model became quite complex, increasing the risk of overfitting. Several strategies were applied to improve performance:

- **Class Imbalance**: **SMOTE** was used to balance the classes, allowing the model to learn more effectively from the underrepresented class.
- **Feature Scaling**: Since Logistic Regression is sensitive to feature scales, **Scaling** was applied to standardize the data, ensuring all features contributed equally.
- **Hyperparameter Tuning**: **GridSearchCV** with **StratifiedKFold** cross-validation was used to optimize key parameters like regularization strength (C) and solver type.
- **Feature Selection**: **RFE (Recursive Feature Elimination)** was applied to remove irrelevant features and reduce the model's complexity.

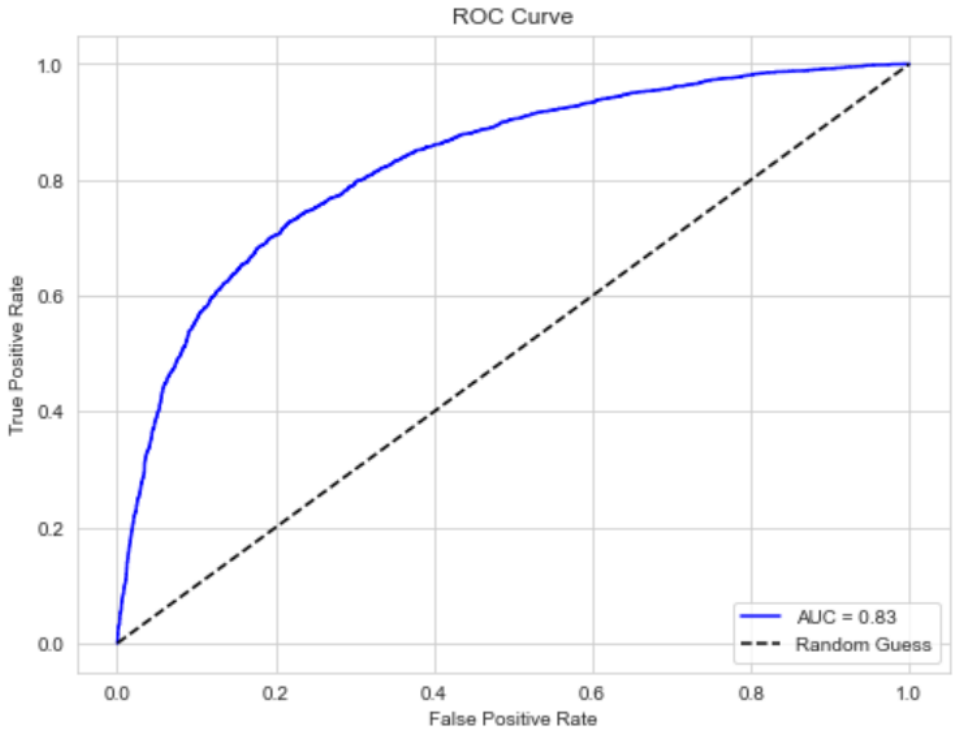
A **Decision Tree** was also used for comparison. Since decision trees do not require scaling, the focus was on hyperparameter tuning to avoid overfitting, adjusting parameters such as max_depth and min_samples_leaf. Like the logistic regression model, **StratifiedKFold** cross-validation was used to maintain class balance.

These optimizations helped prevent overfitting and improved the performance of both models, especially on the large dataset with many features.

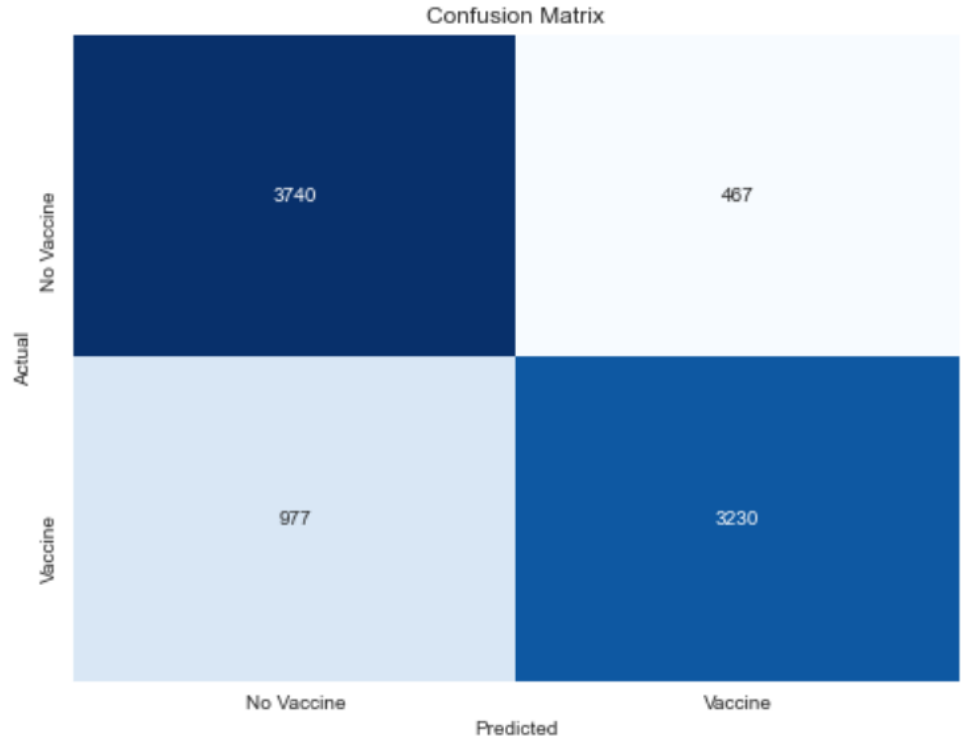
Results

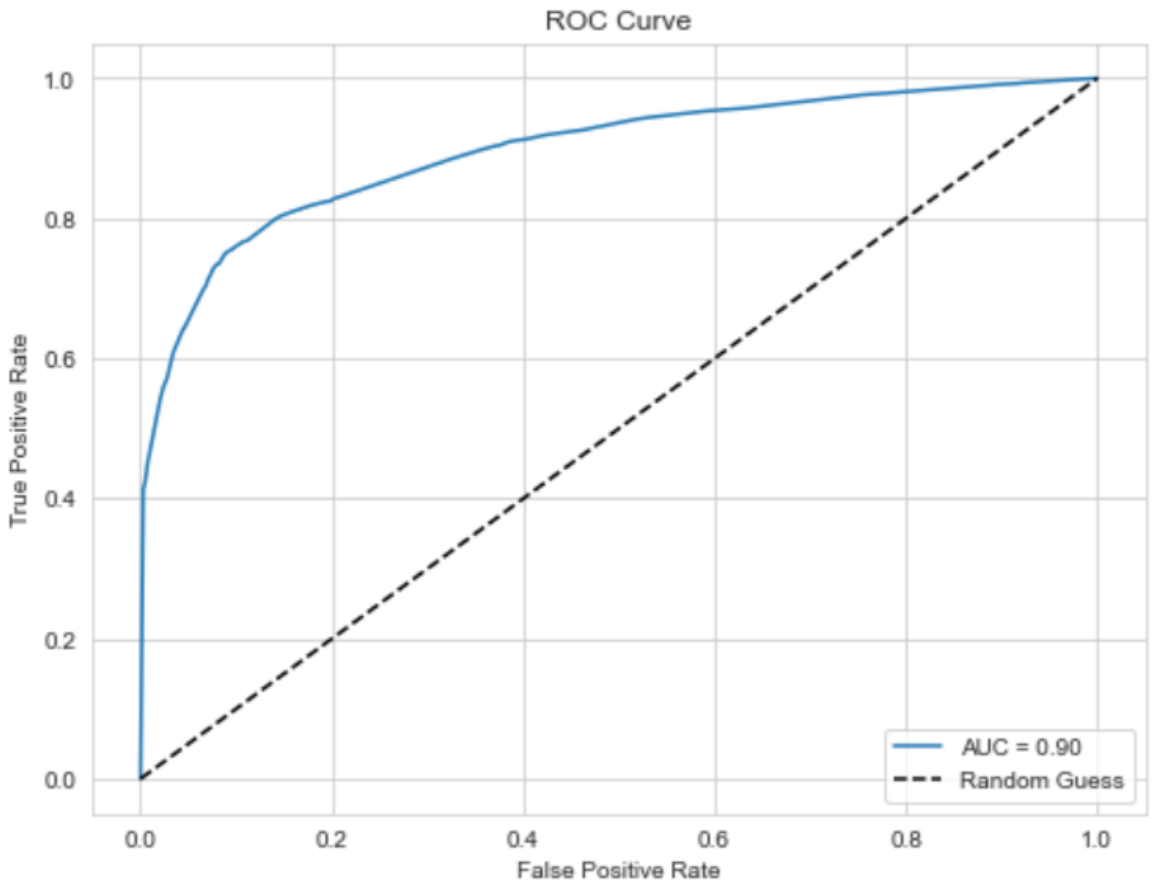
EVALUATION

Confusion matrix and ROC curve – Logistic Regression



Confusion Matrix and ROC curve for Decision Tree





Results vs Metrics

Interpretation of Results vs Metrics

Logistic Regression Model:

- **Best Parameters**:
  - `logreg__C`: 1, indicating regularization strength.
  - `logreg__penalty`: 'l2', the Ridge regularization.
  - `logreg__solver`: 'liblinear', efficient for smaller datasets.

1. **Accuracy (75.49%)**:
   - This score is in the "Good" range (70-80%), suggesting decent performance. However, since the dataset might be imbalanced, accuracy alone is insufficient for evaluating performance.
2. **Precision (0.74 for class 0, 0.77 for class 1)**:
   - Below the desired threshold of 80%. This indicates that when the model predicts a class, it's correct most of the time but still prone to false positives.
3. **Recall (0.78 for class 0, 0.73 for class 1)**:

    - Both values are below 80%, meaning the model is missing some actual positive cases (false negatives). This is a concern, especially for class 1.
4. **F1 Score (0.76 for both classes)**:
    - Falls short of the target (>0.8). It suggests that the balance between precision and recall is moderate but could be improved.
5. **AUC-ROC Score (0.828)**:
    - This score is slightly below the ideal threshold (0.85). The model performs reasonably well in distinguishing between the classes, but there is room for improvement.

Summary: The logistic regression model shows moderate performance but struggles with precision, recall, and overall F1 score. It's effective but not highly reliable for imbalanced datasets.

Decision Tree Model:

- **Best Parameters**:
    - `criterion`: 'entropy', indicating information gain is used for splits.
    - `max_depth`: 10, limiting the depth to prevent overfitting.
    - `min_samples_leaf`: 4, ensuring minimum samples in each leaf node.
    - `min_samples_split`: 10, requiring minimum samples to split a node.

1. **Accuracy (82.84%)**:
    - Above the "Good" range, indicating strong performance overall. This suggests that the decision tree is better suited for this dataset.
2. **Precision (0.79 for class 0, 0.87 for class 1)**:
    - Precision for class 1 exceeds the 80% threshold, showing that predictions for vaccinated individuals (class 1) are more accurate and less prone to false positives.
3. **Recall (0.89 for class 0, 0.77 for class 1)**:
    - Recall for class 1 (vaccinated individuals) is below 80%, meaning the model misses some actual positives. Class 0 recall is very strong, suggesting better coverage for unvaccinated individuals.
4. **F1 Score (0.84 for class 0, 0.82 for class 1)**:
    - Both scores exceed the target (>0.8), indicating a balanced model that handles precision and recall well.
5. **AUC-ROC Score (0.896)**:
    - Exceeds the 0.85 threshold, showing excellent performance in distinguishing between classes, even across various thresholds.

Summary: The decision tree outperforms logistic regression in all key metrics. While its recall for class 1 is slightly below the ideal, its precision, F1 score, and AUC-ROC suggest it is a robust model for this dataset.

CONCLUSION

The primary objective of predicting H1N1 vaccination status using various factors was achieved with significant insights into the effectiveness of two predictive models:

1. **Logistic Regression**:
    - Achieved moderate accuracy (75.49%) and AUC-ROC (0.828), demonstrating reasonable classification performance.
    - However, the model struggled to achieve high precision and recall, especially for predicting vaccinated individuals, indicating challenges with imbalanced data.
2. **Decision Tree**:
    - Outperformed logistic regression with an accuracy of 82.84% and an AUC-ROC of 0.896.
    - Achieved strong precision (0.87 for vaccinated) and recall (0.89 for not vaccinated), making it highly effective at distinguishing between vaccinated and unvaccinated individuals.
    - The results suggest that the decision tree is more suitable for this dataset due to its ability to handle complex feature interactions.

3. RECOMMENDATIONS

1. **Model Selection**:
    - The decision tree should be prioritized for deployment due to its superior performance across key metrics.
2. **Handling Imbalanced Data**:
    - Apply advanced oversampling techniques like SMOTE or consider ensemble methods (e.g., Random Forest, Gradient Boosting) to further improve recall for underrepresented classes.
3. **Feature Importance**:
    - Perform an in-depth analysis of feature importance to identify the key drivers of vaccination status. This can guide public health strategies by focusing on influential factors.
4. **Interpretability**:
    - Decision trees provide clear, interpretable decision rules, making it easier to communicate insights to stakeholders.
5. **Validation**:
    - Conduct further validation on an independent test set or through cross-validation with more folds to confirm the robustness of the results.

6. NEXT STEPS

1. **Model Optimization**:
    - Experiment with ensemble methods like Random Forest and XGBoost to assess if they offer additional gains in performance.

    - Fine-tune hyperparameters of the decision tree further, exploring a larger parameter grid.
2. **Real-World Deployment**:
    - Integrate the decision tree model into a production pipeline, ensuring data preprocessing steps like encoding, scaling, and handling missing values are automated.
3. **Insights for Public Health Policy**:
    - Use model insights to identify factors that strongly correlate with vaccination. Design targeted campaigns addressing these factors to increase vaccination rates.