Predicting H1N1 and Seasonal Flu Vaccine Uptake

Leveraging Data to Improve Public Health

Soudie Okwaro

Data Science Student | Moringa School

A DrivenData Challenge

DRIVENDATA

Introduction

- The Problem: Low vaccination rates lead to increased public health risks.
- The Goal: Predict individual vaccination likelihood for targeted outreach especially targeting H1N1 and Seasonal Flu vaccines.
- Data: National survey data with demographics, health behaviors, and vaccine opinions

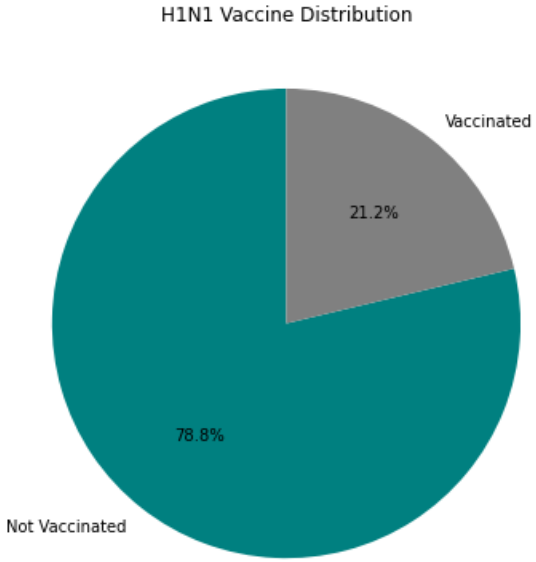
Data Exploration

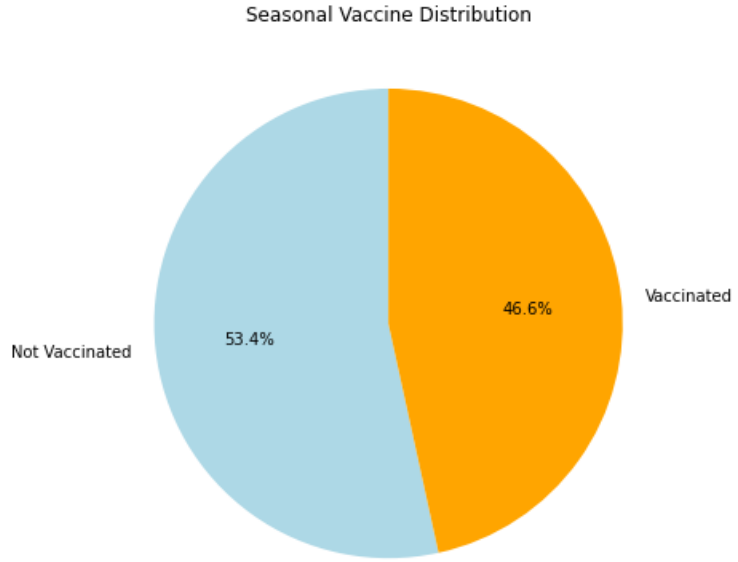
Key Factors:

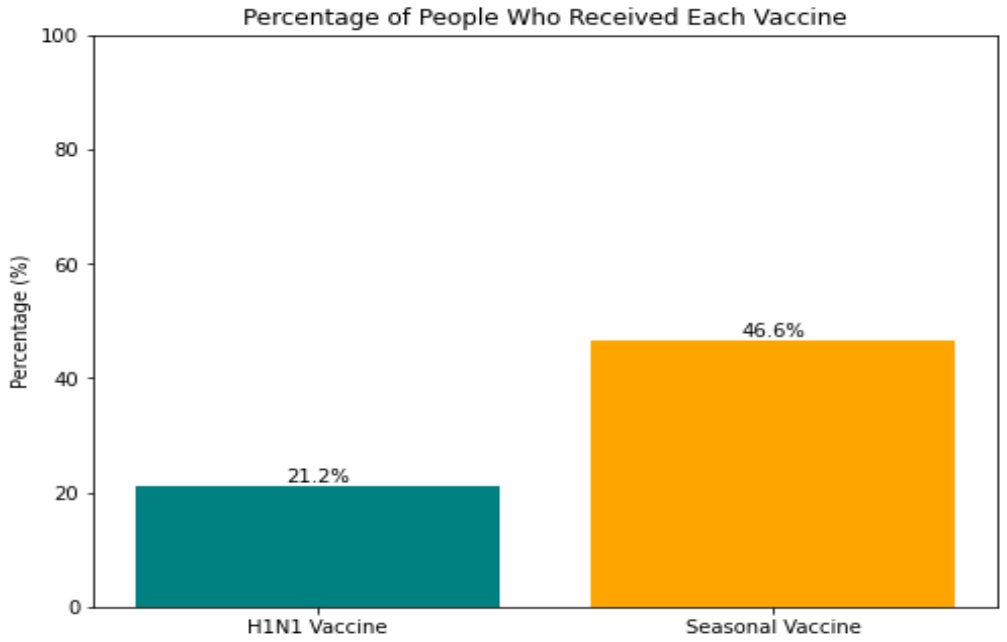
- Behaviors (e.g., hand washing, mask-wearing)
- Beliefs (e.g., vaccine effectiveness, risk perception)
- Demographics (e.g., age, income, education)
- Health Factors (e.g., chronic conditions, health worker status)

Data Imbalance:

More people did not get vaccinated, especially for H1N1.







Modeling Approach

Machine Learning: Models identify patterns in the data to predict who is likely to get vaccinated.

Models Used:

- Logistic Regression: Simple model to understand basic relationships.
- LightGBM: Advanced model to capture complex patterns for better accuracy.

Goal: Select the best model to predict vaccination likelihood.

Model Evaluation

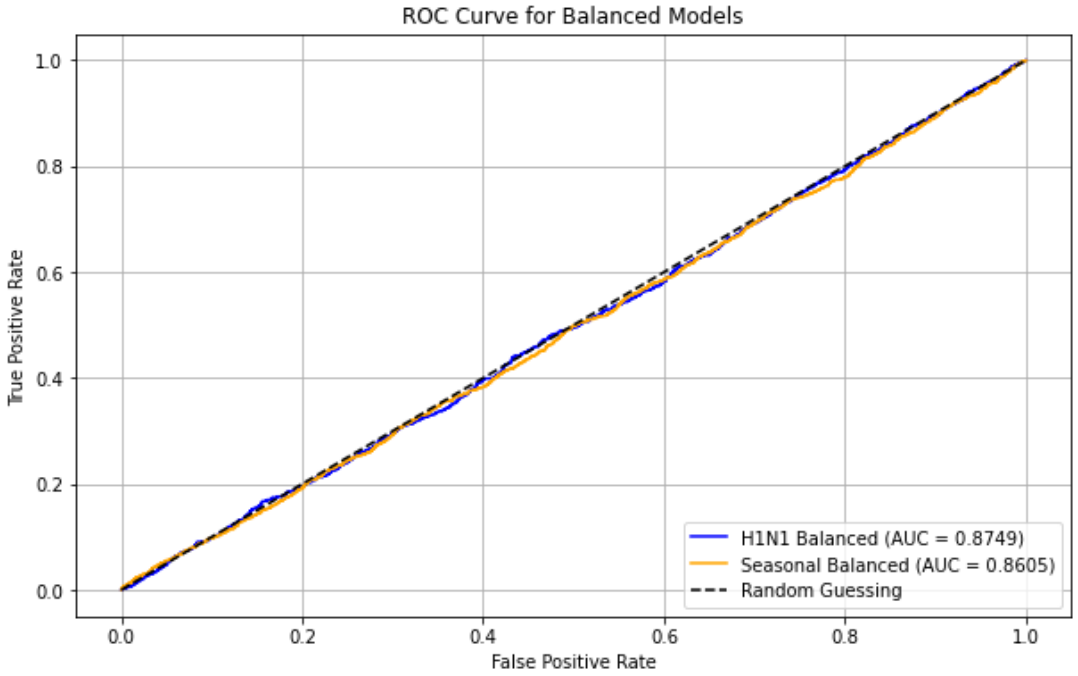
Evaluation Metric:

- AUC (Area Under the Curve) - measures how well the model ranks individuals by their risk of not being vaccinated.

Light GBM Results:

- H1N1 AUC: Around 0.87
- Seasonal AUC: Around 0.86)

The model effectively distinguishes between those more and less likely to get vaccinated.



Results and Impact

Predictions Can Help:

- Identify high-risk groups for targeted outreach.
- Design effective communication strategies.
- Optimize resource allocation for vaccination campaigns.

Potential Benefits:

- Improved vaccination rates, reduced disease spread, and better public health outcomes.

Summary

Business Understanding

(The primary goal was to predict H1N1 and seasonal flu vaccine.

The uptake to improve public health outreach. The motivation was to help public health organizations design effective campaigns and efficiently allocate resources.)

Data Understanding & Preparation

(Data was obtained from the National 2009 H1N1 Flu Survey. Key datasets included training features and labels, test features, and a submission format file. Exploratory Data Analysis (EDA) was performed to understand feature distributions, missing values, and class balance. Target variables (`h1n1_vaccine` and `seasonal_vaccine`) exhibited class imbalance.)

Values in categorical features were imputed with the string 'missing'. Categorical features were encoded using Label Encoding. One-Hot Encoding was explored for the 'education' feature.

Data Modeling

(Logistic Regression and LightGBM models were used. Stratified K-Fold cross-validation was employed to handle class imbalance. Separate models were trained for H1N1 and seasonal vaccine prediction. LightGBM parameters were tuned for optimal performance.)

Evaluation

(AUC (Area Under the ROC Curve) was the primary evaluation metric. Classification reports and confusion matrices were used to assess model performance. LightGBM models outperformed Logistic Regression, demonstrating a stronger ability to predict vaccine uptake.)

Deployment

(The trained LightGBM models were used to predict H1N1 and seasonal flu vaccine uptake on the test set. A submission file (submission.csv) was generated with the predicted probabilities.)

Conclusion

- Project Summary: Successfully predicted flu vaccine uptake using machine learning.
- LightGBM was the best performing model

Thank you!