Lewis Gathua Kamindu

PREDICTING VACCINATION PATTERNS

github.com/lewigi

TABLE OF CONTENT

This document contains the following topics.

- Introduction
- Problem Statement
- Objectives
- Data Understanding
- Models
- Evaluations
- Recommendations
- Conclusions & Next Steps

INTRODUCTION

This work focuses on creating a machine learning project that helps stakeholders in the Public Health Sector predict vaccination patterns for the H1N1 flu and Seasonal flu.

We analyze Backgrounds, Opinions, and Behaviors in H1N1 and Seasonal Flu Vaccination Uptake.

There may be challenges with imbalanced data and ethical concerns but some solutions could be using resampling techniques and ensuring diverse representation in data.

PROBLEM STATEMENT

One of the main SDG'S is Good Health and Well-Being. It is therefore imperative that we use available data to give insights into the betterment of Global Health. This project seeks to leverage data from the National 2009 H1N1 Flu Survey to give actionable insight to relevant stakeholders in Global Health to enhance vaccination strategies, bolster disease prevention, and foster informed decision-making.

OBJECTIVES

Main Objectives

To leverage the data from the National 2009 H1N1 Flu Survey and employ machine learning methodologies to derive actionable insights that contribute to enhancing global vaccination strategies and fostering improved public health outcomes.

Specific Objectives

1. Identify Key Determinants: Analyze the dataset to discern the critical socio-economic, demographic, and behavioral factors influencing individuals' decisions regarding H1N1 and seasonal flu vaccination uptake.

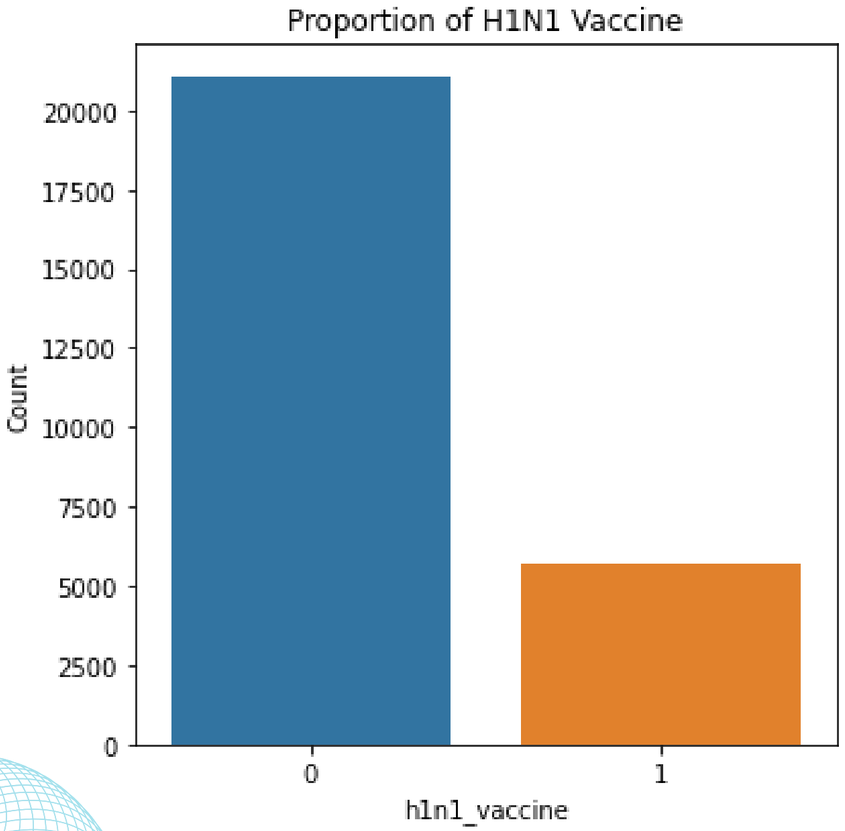
2. Develop Predictive Models: Build, tune and evaluate robust machine learning models to predict and understand the likelihood of individuals receiving the H1N1 and seasonal flu vaccines based on their backgrounds, opinions about illness risks, perceptions of vaccine effectiveness, and health behaviors.

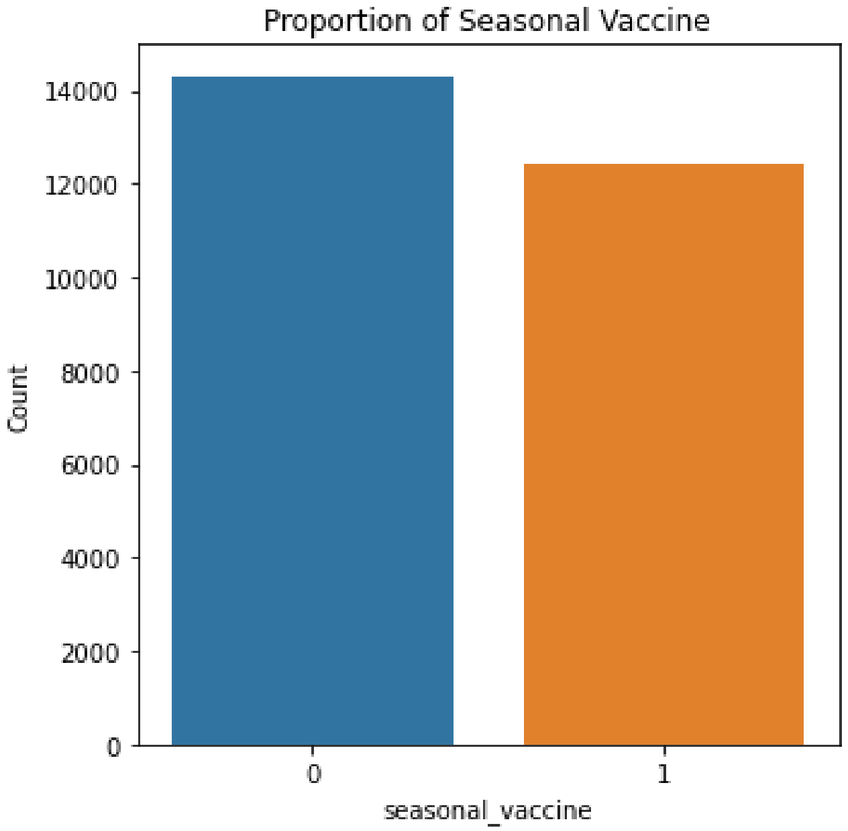
THE DATA

The data has been sourced from U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The National 2009 H1N1 Flu Survey courtesy of the United States National Center for Health Statistics. The data has been divided into train and test features in different excel files. The target columns are h1n1_vaccine and seasonal_vaccine. Some of the 35 available feature columns are:

- h1n1_concern
- h1n1_knowledge
- doctor_recc_h1n1
- health_insurance
- race
- sex
- behavioral_antiviral_meds
- behavioral_face_mask
- chronic_med_condition
- health_worker
- etc

DATA EXPLORATION

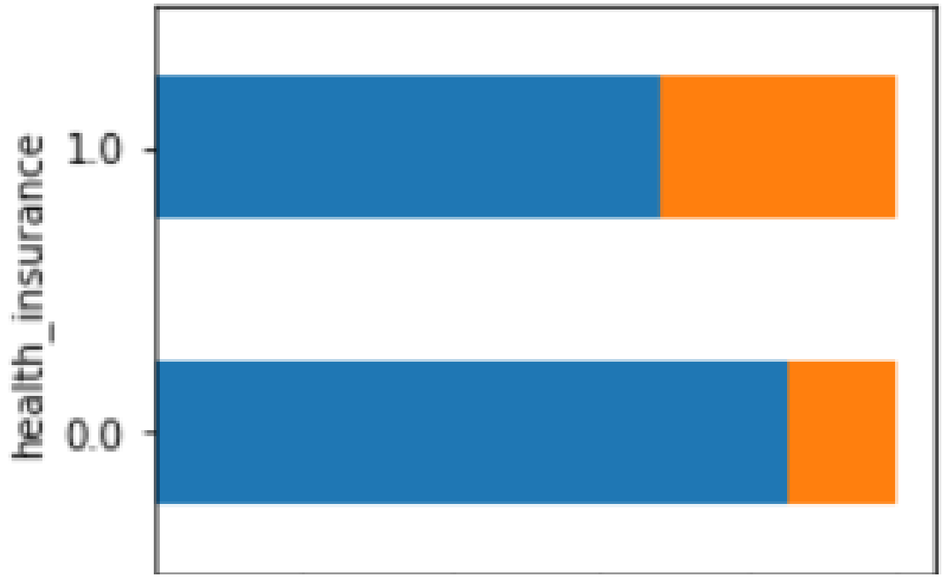


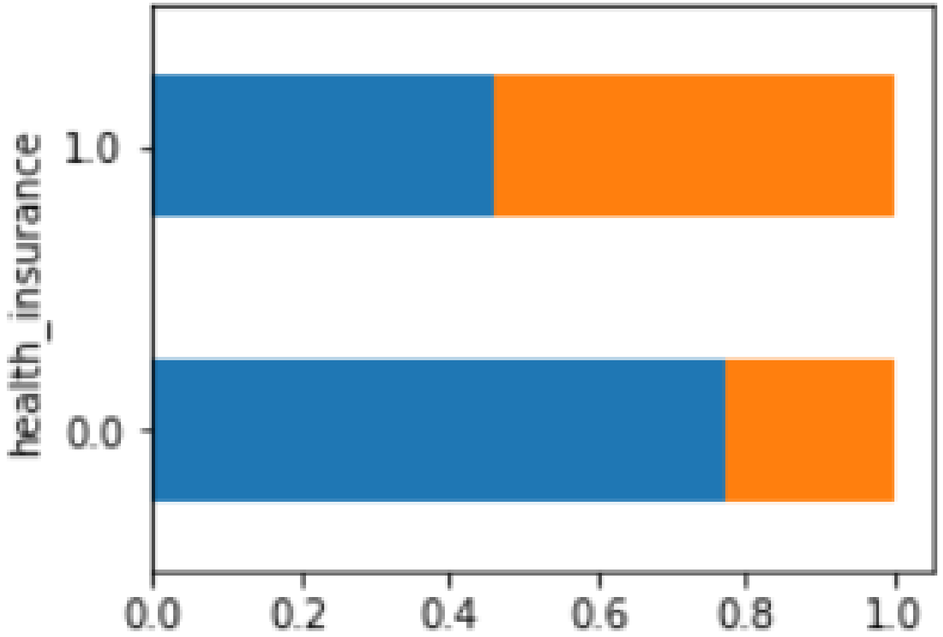


The visualization reveals a significant class imbalance between H1N1 vaccine recipients and non-recipients. To tackle this imbalance in modeling, employing resampling techniques is crucial for class balance. Remarkably, the counts for individuals who received or did not receive the seasonal flu vaccine are nearly equal. Hence, resampling might not be needed for the seasonal flu vaccine variable.

DATA EXPLORATION

Certain features significantly influence the probability of an individual receiving either the H1N1 vaccine or the seasonal vaccine. These influential features include concerns about the virus, knowledge level, opinions regarding vaccine effectiveness, perceptions of H1N1 or seasonal flu risks, recommendations from doctors, and whether one had health insurance as shown in this vizualization.





THE PROCESS

Data preprocessing involved categorical variable handling, feature scaling, encoding, and dataset splitting using Pipelines. Models including Logistic Regression, Random Forest, and K-Nearest Neighbours were developed initially, while XGBoost was utilized for ensemble methods. Further model optimization was pursued through hyperparameter tuning via GridSearchCV. These steps collectively aimed to enhance model performance and accuracy for the given problem domain.

1 Data Preprocessing

Steps were taken to preprocess the data, such as handling categorical variables, feature scaling, encoding, and splitting the dataset into train and test sets with the use of Pipelines.

2 Modelling

A baseline model was developed using Logistic Regression. More complex models were also developed using Random Forest and K-Nearest Neighbours. Ensemble methods were applied using XGBoost.

3 Model Tuning

Further hyperparameter tuning was done using GridSearchCV.

EVALUATION & PERFOMANCE

1. Logistic Regression demonstrated moderate accuracy (66.87%) and a decent ROC AUC score (0.8458).
2. Random Forest, after tuning, improved accuracy (67.03%) and showed a higher ROC AUC score (0.8583).
3. K-Nearest Neighbours, despite tuning, had lower accuracy (60.39%) and a moderate ROC AUC score (0.7663).
4. The XGBoost Classifier performed competitively with a test accuracy of 68.59% and a ROC AUC score of 0.8577.
5. Notably, the XGBoost model fine-tuned with GridSearchCV exhibited the highest accuracy (69.05%) and an excellent ROC AUC score (0.8664), demonstrating superior predictive ability.

The XGBoost model with GridSearchCV stood out as the best-performing model, showcasing the highest accuracy and superior predictive capability among the models evaluated.

RECOMMENDATIONS

1. Tailored Messaging: Use machine learning findings to customize campaign messages, focusing on concerns, knowledge levels, and perceptions to encourage vaccine uptake among specific demographics.
2. Personalized Education: Develop educational programs addressing individual concerns and misconceptions using predictive features identified by machine learning models.
3. Health Insurance Adaptation: Align insurance offerings with policyholder behaviors using insights derived from updated data.
4. Optimized Guidance: Utilize model predictions to equip healthcare providers with targeted information to effectively address patient concerns and increase vaccine acceptance.
5. Continuous Monitoring: Integrate machine learning models into continuous monitoring systems, updating strategies based on evolving trends and factors influencing vaccine uptake.

These data-driven strategies aim to address specific factors affecting vaccine acceptance, potentially leading to improved vaccination rates and public health outcomes.

CONCLUSION & NEXT STEPS

The XGBoost model with GridSearchCV outperformed others, showing the highest test accuracy and ROC AUC score. It demonstrated superior predictive ability and generalization to new data, making it the best-performing model among those evaluated.

Moving forward, several steps can be taken to further enhance model performance and explore new avenues:

1. Deployment and Monitoring: Prepare the top-performing model for deployment and establish monitoring mechanisms to track its real-world performance.
2. Hyperparameter Tuning: Continuously fine-tune hyperparameters to optimize model performance, potentially exploring wider parameter ranges.

THANK YOU !