Problem Statement: Predicting Probabilities of H1N1 and Seasonal Vaccination Uptake.

Abstract

In the face of public health challenges posed by influenza outbreaks, understanding factors influencing individual vaccination probabilities is paramount for effective disease prevention strategies. This report addresses the challenge of predicting individual vaccination probabilities for both H1N1 and seasonal flu separately, employing data from the 2009 National H1N1 Flu Survey (NHFS). We utilize two BaggingClassifiers, each with an XGBoostClassifier as the base estimator, addressing the multilabel nature of the prediction task. By leveraging demographic, socioeconomic, and health-related features, our analysis develops models specifically for estimating H1N1 and seasonal flu vaccination probabilities. These findings provide valuable insights into vaccination decision-making processes and contribute to enhancing public health interventions aimed at promoting influenza vaccination.

1 Introduction

Influenza, a highly contagious respiratory illness, poses a significant threat to public health, particularly for vulnerable populations. While vaccination remains the most effective preventive measure, low uptake rates persist. Understanding factors influencing vaccine decisions is crucial for designing targeted interventions. This study employs machine learning to predict H1N1 and seasonal influenza vaccination probabilities separately, using a BaggingClassifier with XGBoostClassifier for each target variable. We address class imbalance and utilize ROC AUC for evaluation within each model. Additionally, we calculate an overall ROC AUC by combining both models' predictions, providing a comprehensive assessment of performance. By identifying key predictors and focusing on robust ROC AUC scores, this research aims to inform public health strategies and improve vaccine uptake rates.

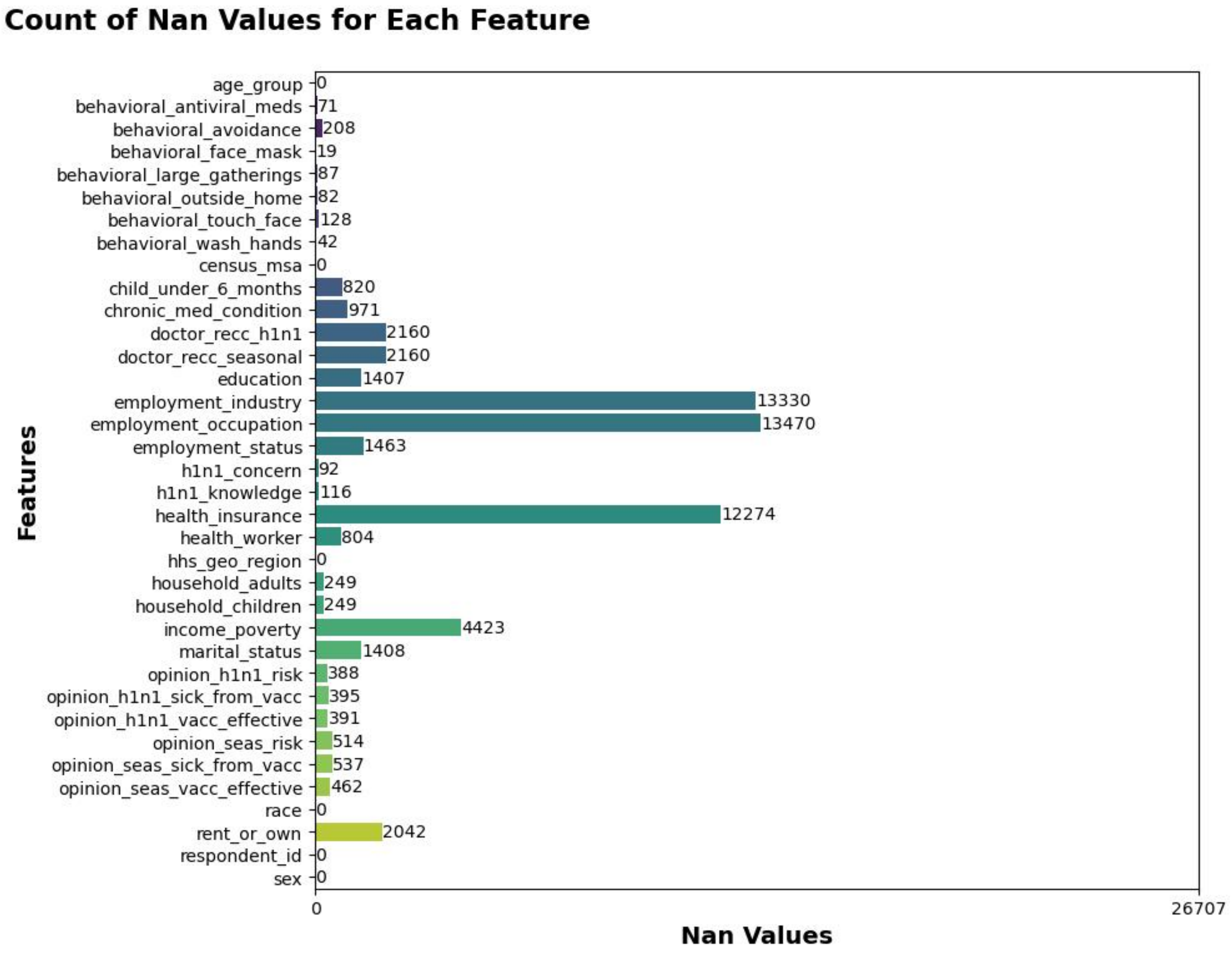
Keywords: BaggingClassifier, XGBoostClassifier, class imbalance, ensemble, h1n1

vaccine, seasonal vaccine, ROC AUC.

2 Exploratory Data Analysis (EDA)

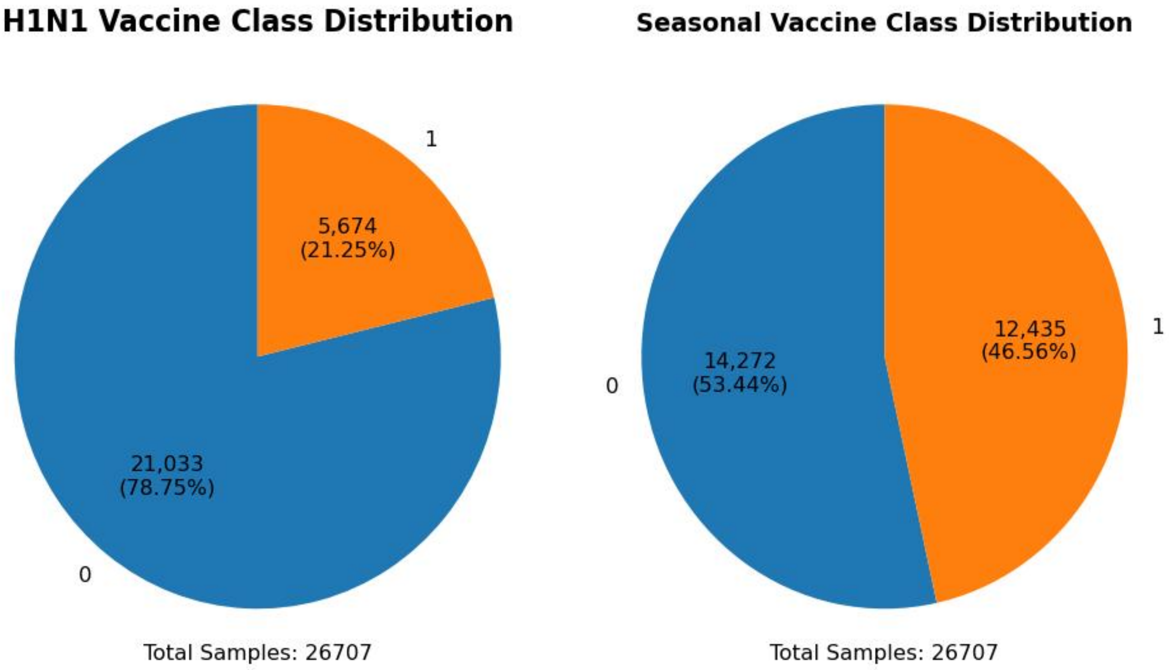
2.1 Null Value Analysis

To gain insights into the presence of missing values within the dataset, we visualized the null values for each feature using bar charts. We identified features with significant missing values, such as ‘health_insurance’, ‘employment_industry’, and ‘employment_occupation’.



2.2 Class Imbalance Analysis

We also investigated the class distribution of the target variables (‘h1n1_vaccine’ and ‘seasonal_vaccine’) to understand the imbalance between different classes. The class imbalance was visualized using pie charts, illustrating the proportion of samples in each class.



3 Dataset Preparation

3.1 Dataset Cleaning

Since the column ‘respondent_id’ is just useful as an index and has no predictive significance, it is eliminated. The columns ‘health_insurance’, ‘employment_industry’, and ‘employment_occupation’ are also dropped due to a substantial amount of missing data in those fields.

3.2 Dataset Preprocessing

Numerical Features: We imputed missing values in numerical features using IterativeImputer. StandardScaler was applied to scale numerical features.

Categorical Features: Missing values in categorical features were imputed using the most frequent strategy. One-hot encoding was performed on categorical features to convert them into a numerical representation.

4 Proposed Methodology

4.1 Addressing Class Imbalance

We calculated 'scale_pos_weight' parameter of XGBoostClassifier based on the class distribution to address class imbalance in the target variables. This approach addresses class imbalance in the target variables, ensuring balanced learning during model training.

$$scale\_pos\_weight = \frac{(len(y\_data) - y\_data.sum())}{y\_data.sum()} = \frac{total\ number\ of\ negative\ samples}{total\ number\ of\ positive\ samples}$$

$$where,\ y\_data\ is\ a\ target\ label$$

4.2 Model Creation and Initialization

We created two XGBoostClassifiers each tailored for one target variable. We then set logistic regression as the objective function for binary classification, with AUC used as the evaluation metric. Additionally, we employed Bagging Classifiers with XGBoostClassifier as the base estimator to harness ensemble learning benefits.

4.3 Fine-tuning Hyperparameters

We utilized Stratified RandomizedSearchCV and Stratified GridSearchCV to explore a comprehensive range of hyperparameters for each model, prioritizing optimization based on the ROC AUC metric. The best hyperparameters found were then used.

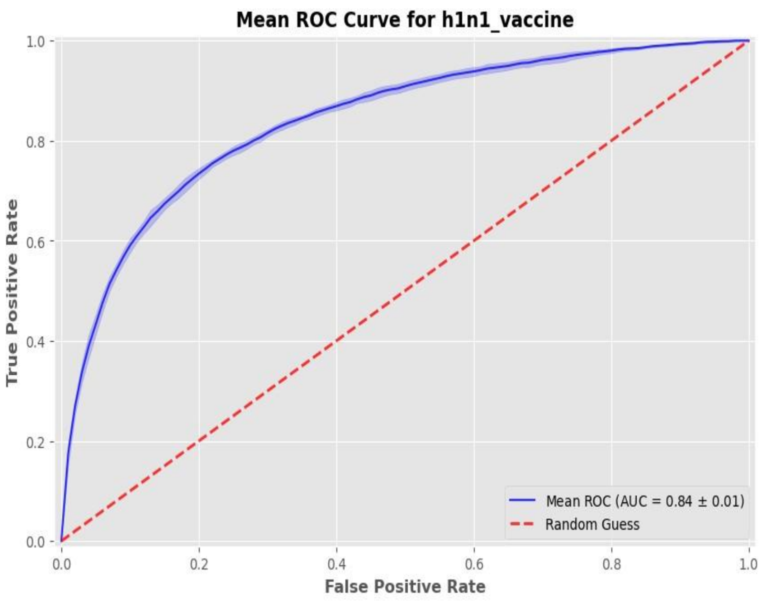
4.4 Model Training and Validation

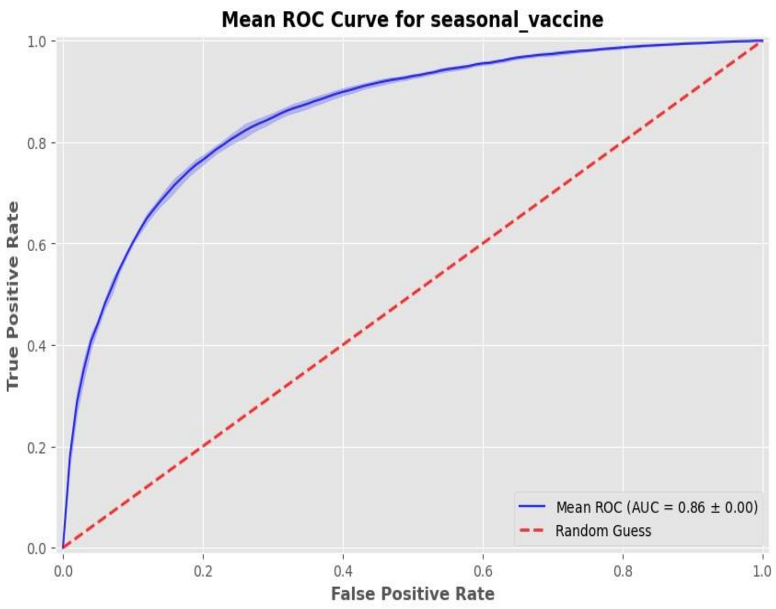
We employed a Stratified K-fold Cross-Validator to split the data into training and testing sets. We then trained and validated each model for each fold, computing the ROC AUC score on the test set. Finally, we calculated the mean ROC AUC score across all folds to assess the overall performance of each model.

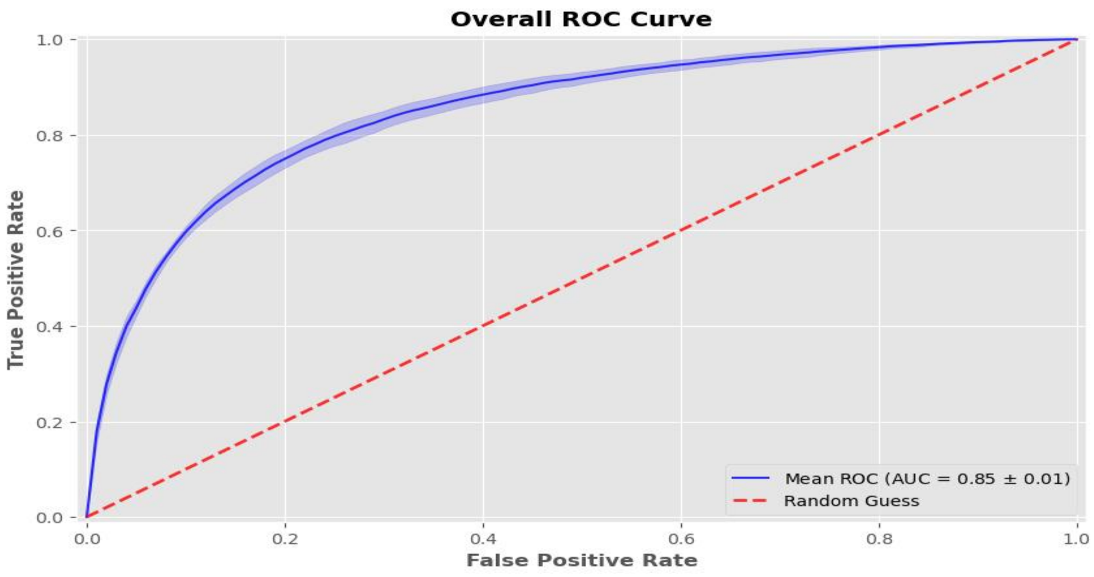
5 Evaluation

The mean ROC AUC scores for each target variable and the overall score are summarized below:

- H1N1 vaccine: Mean ROC AUC score = 0.8415
- Seasonal vaccine: Mean ROC AUC score = 0.8588
- Overall ROC AUC score = 0.8501







While employing separate models and predicting probabilities for each target variable (h1n1 and seasonal vaccine), the narrow confidence intervals and low standard deviations observed in the visualizations of both individual and combined AUC scores across folds demonstrate consistent and reliable performance. This suggests stability in the models' ability to discriminate between positive and negative samples for both target variables.

6 Conclusion

In conclusion, our study employed BaggingClassifier with XGBoostClassifier as base estimator to predict H1N1 and seasonal flu vaccination probabilities from the 2009 NHFS data, achieving ROC AUC scores of 0.8415 (H1N1 vaccine), 0.8588 (seasonal vaccine), and 0.8501 (overall). These findings offer valuable insights for public health officials. By understanding the factors associated with vaccine uptake, public health efforts can be strategically targeted to specific populations and tailored to address the most influential decision-making factors. This ultimately contributes to improved vaccine uptake rates, enhanced public health preparedness, and mitigation of the impact of influenza outbreaks.