data_report

September 12, 2025

1 Project Data Report: Predicting Uptake for H1N1 Vaccine Using Machine Learning

1.1 OVERVIEW

This study uses data from the National 2009 H1N1 Flu Survey (NHFS) to examine vaccination uptake during the 2009 H1N1 influenza pandemic. Developing predictive models that identify the variables affecting vaccination choices for the seasonal flu and H1N1 vaccines is the aim. Our goal is to improve vaccination rates and stop future outbreaks by using machine learning techniques to provide insights for public health strategies.

The dataset comprises 26,707 respondents that cover perceptions, health behaviors, and demographics. These are the main findings. A low overall uptake rate was found by exploratory analysis (21% for H1N1 and 36% for seasonal flu), with correlations found to variables such as health concerns and physician recommendations. Features like "doctor_recc_h1n1" were highlighted as powerful predictors by the models (logistic regression, decision trees, and random forests), which achieved a respectable accuracy (~85% for H1N1 and ~76% for seasonal). Recommendations include educating people about vaccine hesitancy and launching targeted campaigns for high-risk populations.

The dataset's age (2009), which might not accurately reflect current behaviors, and possible survey bias are among its limitations. New data or sophisticated models like neural networks may be used in future research.

This study illustrates how data-driven strategies can improve public health readiness and possibly minimize the effects of infectious diseases.

1.2 Introduction

2 Business Problem

A vital component of public health, vaccinations save lives and stop the spread of infectious diseases. Vaccines were created and made available to lessen the effects of the 2009 H1N1 ("swine flu") pandemic, which is estimated to have killed between 151,000 and 575,000 people worldwide in its first year. But because of things like misinformation, hesitancy, and access problems, uptake was poor.

What factors influence vaccination uptake for seasonal flu and H1N1 and how can we predict it to inform targeted interventions? This project answers this question.** Through the analysis of NHFS survey data, we create predictive models to pinpoint at-risk populations and suggest tactics for organizations, legislators, and medical professionals to increase vaccination

rates. For future pandemics, where quick, widespread adoption can lower mortality and financial expenses, this is essential.

3 Objectives

- Comprehend the dataset and determine the key factors impacting vaccination usage.
- Clean up and prepare data for modeling.
- To find patterns, do exploratory data analysis (EDA).
- Engineer features to enhance the performance of the model.
- Create and assess machine learning models to forecast vaccination uptake.
- Make practical suggestions based on the insights from the model.

4 Scope and Limitations

- Scope: Use survey data to focus on binary classification for H1N1.
- Restrictions: The data is U.S.-centric (may not generalize globally), self-reported (possible bias), and from 2009.

5 Stakeholder Analysis

Stakeholders consist of: - **Public Health Officials and Governments**: Utilize models to plan campaigns and distribute resources (e.g., focus on low-uptake areas). **Healthcare Providers**: Determine patient profiles for tailored advice (e.g., give priority to high-risk groups such as individuals with chronic conditions). - **NGOs and policymakers**: Reduce hesitancy by informing policies on vaccine access and education. The general public indirectly benefits from better vaccination practices, which result in herd immunity.

Stakeholder prioritization: For immediate impact, concentrate on officials and providers, as models can direct interventions.

5.1 Data Understanding

6 Data Sources

Source: CDC's National 2009 H1N1 Flu Survey (NHFS)

Records: 26,707 U.S. citizen survey responses.

h1n1_vaccine is the target variable (1 = vaccinated, 0 = not vaccinated).

Qualities:

Opinions and Knowledge: perceived efficacy, knowledge, and concern.

Behaviors include wearing masks, washing your hands, and staying away from crowds.

Age, income, race, insurance, and region are examples of demographics.

External influences include medical advice and employment in the healthcare industry.

Relevance: Offers rich predictors for evaluating vaccination choices, locating vulnerable populations, and guiding public health initiatives.

6.1 Data Preparation and Cleaning

7 Handling Missing Values

- Determined which columns (such as `health_insurance` and `employment_industry`) had more than 40% missing.
- Imputation: Used median for numerical and mode for categorical .
- dropped employment_industry and employment_occupation columns which were missing over 13,000 entries around 50% of the data

8 Encoding Categorical Variables

Using stratified sampling, we first divided the dataset into training (80%) and testing (20%) sets, making sure that the proportion of vaccinated versus non-vaccinated people in each set was balanced. This was crucial since our target variable (h1n1_vaccine) is extremely imbalanced, and if stratification isn't used, one set may have too few vaccinated cases, which could result in inaccurate training or assessment.

- We fixed missing values and made sure the formatting was consistent for preprocessing:
- Numerical columns: StandardScaler was used to scale features so that all values fall within a similar range, and the median was used to fill in missing values.
- Categorical columns: We used One-Hot Encoding to turn categories into binary (0/1) columns and filled in missing values with the most common category (mode).

We were able to apply preprocessing consistently throughout model training and testing by combining these steps into a single ColumnTransformer.

9 Encoding Categorical Variables

10 EDA after data preparation

Using a heatmap, we looked at feature correlations. The findings indicated that there were no pairs of variables with extremely high collinearity ($|r| > 0.8$) and that the majority of features had low to moderate correlations.

This suggests that multicollinearity is not an issue for this dataset. As a result, we can keep every feature for modeling without deleting or combining variables.

11 Class Imbalance Check

- There is a significant imbalance in the target variable (H1N1 vaccine) bar chart:
- The vast majority of respondents (class 0) did not receive a vaccination.
- Only a smaller percentage (class 1) were vaccinated.

- Because of this imbalance, a model could achieve high accuracy by primarily predicting the majority class, making accuracy by itself deceptive.

We will evaluate this using metrics like F1-score, AUC-ROC, Precision, and Recall. To address this imbalance during modeling, we might also take into account class weighting or resampling strategies.

12 Data Preparation Conclusion

- We have finished all the procedures required to get the dataset ready for modeling:
- developed fresh, more educational features (prevention index, opinion scores, etc.).
- Features that could lead to data leakage, duplicate columns, and dropped IDs.
- To maintain class balance, divide the data into training (80%) and testing (20%) sets using stratification.
- One-hot encoding for categorical features, scaling for numerical features, and imputation for missing values are examples of applied preprocessing.
- confirmed that multicollinearity would not be caused by any highly correlated features.
- verified that the h1n1 vaccine target variable is unbalanced and needs extra attention when being evaluated.

At this point, the dataset is completely numeric, clean, and prepared for modeling.

12.1 Modeling & Evaluation

- We will use an iterative approach to our modeling strategy:
- Start with a baseline model that is easy to understand.
- Test increasingly intricate models one after the other.
- To enhance performance, use hyperparameter tuning.
- To guarantee dependability and prevent overfitting, cross-validation will be used to validate every model.
- Recall for the positive class (vaccinated = 1) will be our main evaluation metric. False negatives are more expensive than false positives in this public health setting. It is more dangerous to miss someone who would genuinely get vaccinated (false negative) than to make a false positive prediction.
- This guarantees that the finished model is accurate and in line with the practical objective of identifying the greatest number of possible vaccine recipients for focused interventions.

13 Cross-Validation Results

- 5-fold cross-validation was used to assess the baseline Logistic Regression model, with recall serving as the main metric:
- Scores for CV Recall: [0.66, 0.71, 0.68, 0.68, 0.68]

- Mean CV Recall: 0.68
- 0.017 is the CV Recall Standard.

This indicates that, with minimal variation across folds, the model accurately identifies roughly two-thirds of vaccinated individuals. Although this is a good place to start, it also shows that about one in three vaccinated people are overlooked, underscoring the need for a more robust model to reduce false negatives.

14 Training vs. Test Evaluation

The model was then fitted on the entire training set and evaluated on both training and test sets:

- Training Recall (vaccinated=1): 0.69
- Test Recall (vaccinated=1): 0.68
- Training Accuracy: 0.77
- Test Accuracy: 0.76

The close similarity between training and test performance suggests the model is not overfitting and generalizes well to unseen data. This stability is likely due to the use of class weighting and the simplicity of logistic regression.

15 Model Interpretation

Performance: 68% of vaccine recipients are identified by the model, which has a test recall of 0.68. For class 1, however, its precision is comparatively low (0.46), indicating that almost half of the positive predictions are wrong. Given that only about 21% of respondents were vaccinated, this imbalance is to be expected.

Overfitting Check: Generalization is confirmed when there is little variation between training and test results.

Implications for Business:

Strength: By capturing the majority of the positive class, the model offers a trustworthy baseline for identifying vaccinated individuals.

Limitation: Because of the low precision, resources might be used to target people who are unlikely to receive vaccinations.

Acceptable Trade-off: False negatives are more expensive than false positives in the context of public health. The effectiveness of the intervention is compromised when someone who would be vaccinated is missed, and the cost is only increased when more people are targeted.

16 Random Forest

Random Forest – Model Evaluation Performance Interpretation

- Training Recall (vaccinated=1): 0.99
- Test Recall (vaccinated=1): 0.36

- Test Accuracy: 0.82
- The Random Forest model shows severe overfitting. While it almost perfectly identifies vaccinated individuals in training, on the test set it only captures 1 in 3 cases. Accuracy is misleading here, as it mainly reflects strong performance on the majority class (not vaccinated).
- Overfitting Analysis
- Recall (train vs. test): 0.99 → 0.36
- Accuracy (train vs. test): 0.99 → 0.82
- This large performance gap confirms that the model memorized training data but fails to generalize.
- Business Implications

A recall of 0.36 for vaccinated individuals means the model misses nearly two-thirds of true positives, making it unsuitable for public health objectives. While its precision is relatively higher (0.62), the low recall undermines its usefulness for identifying at-risk populations.

17 Logistic Regression with Hyperparameter Tuning

- GridSearchCV hyperparameter tuning of logistic regression did not significantly enhance performance over the baseline. For vaccinated individuals, accuracy slightly declined (0.78 vs. 0.76), precision remained steady (0.46 vs. 0.45), and recall remained nearly unchanged (0.68 baseline vs. 0.67 tuned). This suggests that tuning did not improve recall because the baseline model was already nearly ideal.
- Modifying the classification threshold had a greater effect. Recall significantly improved from 0.68 to 0.77 when the threshold was lowered from 0.5 to 0.4, enabling the model to accurately identify 77% of vaccinated people. The anticipated trade-off of adding more false positives was reflected in the precision dropping to 0.38.
- With nearly identical training and test performance (recall: 0.78 vs. 0.77, accuracy: 0.73 vs. 0.72), the model is still demonstrating strong generalization. This demonstrates that recall was enhanced by the threshold adjustment without causing overfitting.
- This change is very helpful from a business point of view. More than three-quarters of the vaccinated population can now be identified by the model, greatly increasing the coverage for public health targeting. This is a reasonable price to ensure that the majority of actual vaccine recipients are enrolled, even though some inefficiency is introduced (6 out of 10 flagged individuals will not vaccinate).

17.1 Final Model Evaluation & Conclusion

Implications - 77% of vaccinated people can be identified using the optimized logistic regression model, which makes it useful for focusing public health messaging. It can be used by agencies to create customized campaigns and allocate resources effectively. Outreach efforts can focus on people who are unlikely to get vaccinated.

Restrictions - Results may not be generalizable because the model is based on survey data from a particular time and place. Doctor recommendations are one example of a feature that may not always be available. Outreach must reach more people than is necessary due to a high false positive rate of 71%.

Suggestions - Target audiences with high doctor recommendations and opinion scores. Gather data on vaccine attitudes in real time. Before scaling, test the model regionally.

Next Actions - To increase accuracy without overfitting, test advanced models like XGBoost and investigate new characteristics.