H1N1 VACCINATION SHOT PREDICTION USING MACHINE LEARNING

Nikhil Kaundal

Abstract

In the wake of the recurring threat posed by the COVID-19 virus, timely and accurate prediction of vaccination shot requirements is crucial for public health preparedness.

This study proposes a machine learning-based approach to predict the demand for H1N1 vaccination shots.

Leveraging historical vaccination data, demographic information, and behavioral factors, a predictive model is developed to forecast the anticipated uptake of H1N1 vaccines within a specified population.

The methodology integrates various machine learning algorithms, including decision trees, random forests, naive bayes and gradient boosting, to analyze past vaccination trends and identify key predictors influencing vaccination rates.

Feature engineering techniques are employed to extract relevant features from the data source, enhancing the model's predictive performance.

Validation of the predictive model is conducted using performance metrics such as accuracy, precision, recall, F1-score and receiver operating characteristic (ROC) analysis.

Problem Description

The **aim of this study** is perform a classification of how likely individuals are to receive their H1N1 flu vaccine.

The classification is going to be done based upon the behavior, opinion , concern and knowledge relating to the H1N1 influenza virus.

The prediction output of this study will give public health professionals and policy makers a clear understanding of factors associated with low vaccination rates.

This in turn enables them to systematically act on those features hindering people to get vaccinated.

Data Overview

In late 2009 and early 2010, the United States conducted the National 2009 H1N1 Flu Survey. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, in conjunction with questions about themselves.

These additional questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviours towards mitigating transmission.

Data source: This Data was collected from Centers for Disease Control and Prevention (National Public Health Agency of the United States).

https://www.cdc.gov/nchs/nis/data_files_h1n1.htm

The data inquires about whether or not people received the the H1N1 flu vaccination, as well as their demographic, behavioral, and health factors.

A total of number of 70,944 responses are listed in this dataset.

Methodology

As the data contains 171 columns but all are not needed for prediction , so the features related to their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission were included.

After dropping the rest of the features ,39 features were left for model building.

Data Cleaning:

- Checking Null Values:
  - Initial examination of the dataset revealed some columns having more than 20 % missing values.
- Dropping Rows having Null Values:
  - Rows from the columns “CONCERN_DKNW_F', 'CONCERN_NONE_F', 'CONCERN_NOTV_F’, 'CONCERN_REFD_F', 'CONCERN_SOME_F',

Methodology

  - 'CONCERN_VERY_F', 'KNOW_H1N1_ALOT_F', 'KNOW_H1N1_DKNW_F', 'KNOW_H1N1_LITL_F', 'KNOW_H1N1_NONE_F', 'KNOW_H1N1_REFD_F' "having missing values were removed

- Dropping Columns:
  - Columns "INSURE', 'Q95_INDSTR' , 'INC_CAT1" having High percentage of missing values were removed.
- Imputing Missing Values:
  - The missing values in the reset of the columns were imputed using mode for categorical features and median for numerical features.
- The dataset after data cleaning has 56,656 entries and 36 features.

Feature Engineering:

New feature "target variable " was created from Vacc_H1N1_COUNT column where count of more than 1 was take as 1 and count of 0 was taken as 0.

Methodology

  - 'CONCERN_VERY_F', 'KNOW_H1N1_ALOT_F', 'KNOW_H1N1_DKNW_F', 'KNOW_H1N1_LITL_F', 'KNOW_H1N1_NONE_F', 'KNOW_H1N1_REFD_F' "having missing values were removed

- Dropping Columns:
  - Columns "INSURE', 'Q95_INDSTR' , 'INC_CAT1" having High percentage of missing values were removed.
- Imputing Missing Values:
  - The missing values in the reset of the columns were imputed using mode for categorical features and median for numerical features.
- The dataset after data cleaning has 56,656 entries and 36 features.

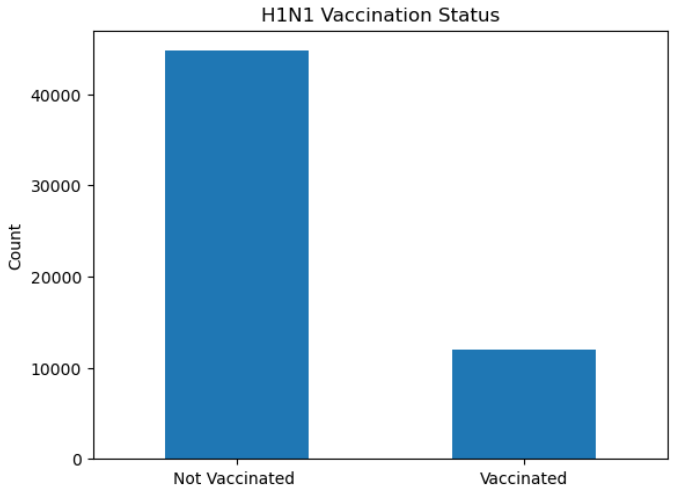
Feature Engineering:

New feature "target variable " was created from Vacc_H1N1_COUNT column where count of more than 1 was take as 1 and count of 0 was taken as 0.

Methodology

Exploratory Data Analysis

- **Vaccination Status**

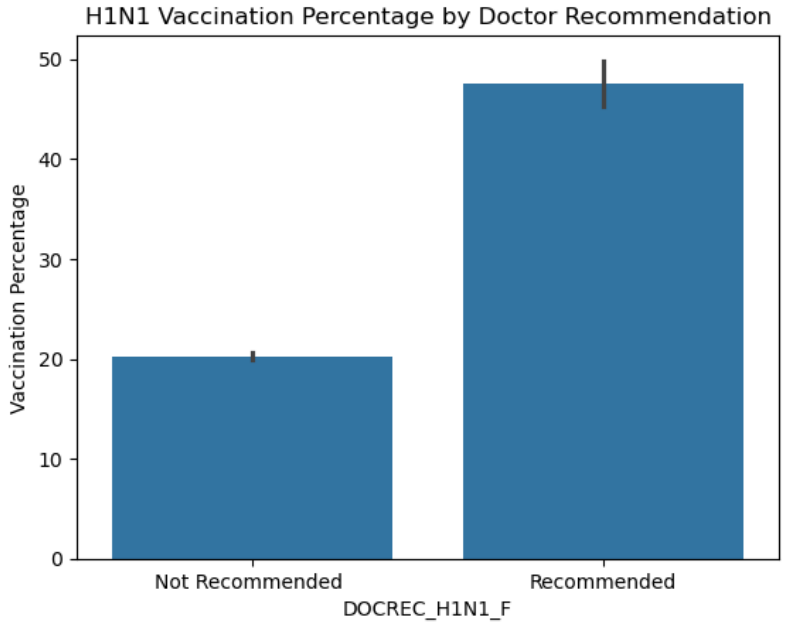


Only **21**% have taken the H1N1 Vaccine

Methodology

Exploratory Data Analysis

- **Relationship vs Doctor recommendation and H1N1 vaccination status**



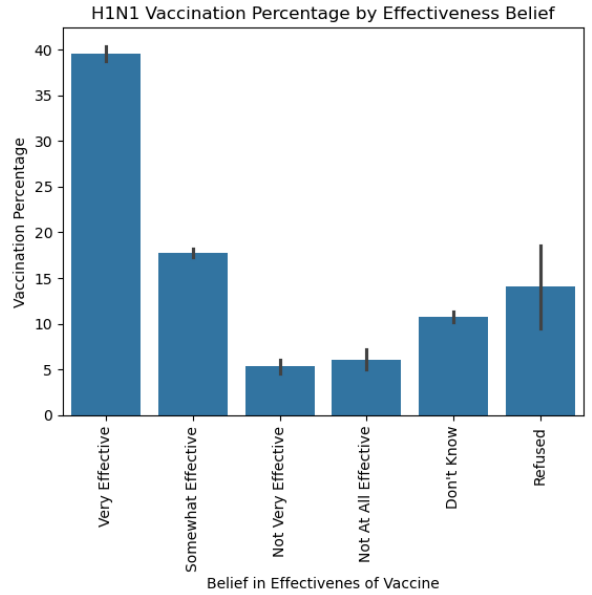
Those who were ***recommended*** by the doctor to take the vaccine of them **50%** have taken the vaccine.

Those who were ***not recommended*** by the doctor to take the vaccine of them only **20%** have taken the vaccine.

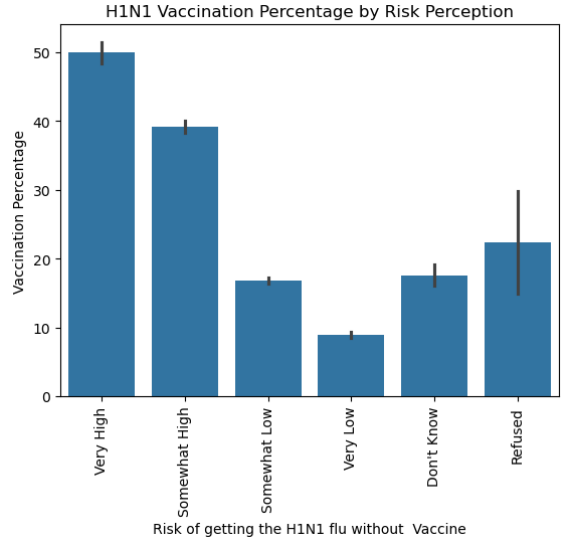
Methodology

Exploratory Data Analysis

- **H1N1 vaccination by effectiveness belief**
  - The percentage of vaccination increase with the increase in the belief in the effectiveness of the H1N1 vaccine.



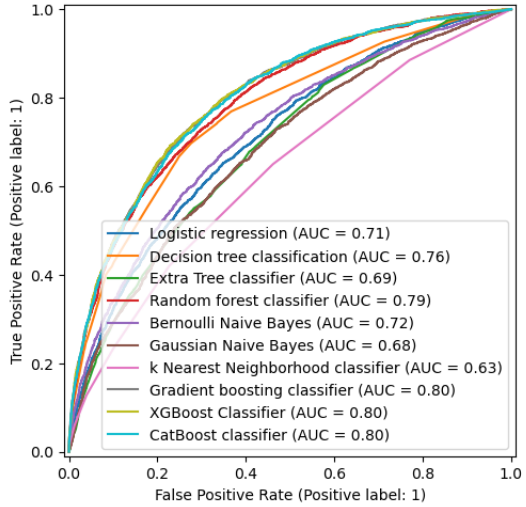
- **H1N1 vaccination by risk perception**
  - The percentage of vaccination increase with the increase in the perception of getting sick with H1N1 flu without vaccine.



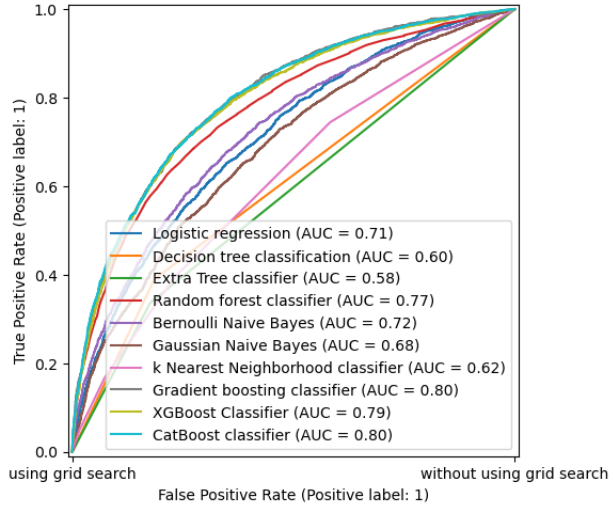
Methodology

Model Evaluation

- The training and testing data was split in the ratio of 80:20.
- A variety of different models were applied to find the most accurate model.
- First the accuracy score, precision score, f1 score, recall and roc-auc curve was examined for each model without any hyperparameters.
- After that hyperparameters were tunned.
- To find the hyperparameters for each model GridSearchCV was used to find the best combinations for each model.



Roc curves with hypertunning



Roc curves without hypertunning

Results

Before Hyperparameter Tunning

1. Logistic Regression
   - Test accuracy 0.7962
   - ROC-AUC score 0.71
2. Decision Tree
   - Test accuracy 0.7253
   - ROC-AUC score 0.60
3. Extra tree classifier
   - Test accuracy 0.7214
   - ROC-AUC score 0.58
4. Random Forest classifier
   - Test accuracy 0.8121
   - ROC-AUC score 0.77
5. Bernoulli Naive Bayes
   - Test accuracy 0.7796
   - ROC-AUC score 0.72

After Hyperparameter Tunning

1. Logistic Regression
   - Test accuracy 0.7962
   - ROC-AUC score 0.71
2. Decision Tree
   - Test accuracy 0.8108
   - ROC-AUC score 0.76
3. Extra tree classifier
   - Test accuracy 0.7910
   - ROC-AUC score 0.69
4. Random Forest classifier
   - Test accuracy 0.8162
   - ROC-AUC score 0.79
5. Bernoulli Naive Bayes
   - Test accuracy 0.7796
   - ROC-AUC score 0.72

Results

Before Hyperparameter Tunning

6. Gaussian Naive Bayes

Test accuracy	0.7135
ROC-AUC score	0.68

7. KNN classifier

Test accuracy	0.7672
ROC-AUC score	0.62

8. Gradient boosting classifier

Test accuracy	0.8179
ROC-AUC score	0.80

9. XG Boost classifier

Test accuracy	0.8159
ROC-AUC score	0.79

10. Cat Boost classifier

Test accuracy	0.8215
ROC-AUC score	0.80

After Hyperparameter Tunning

6. Gaussian Naive Bayes

Test accuracy	0.7135
ROC-AUC score	0.68

7. KNN classifier

Test accuracy	0.7852
ROC-AUC score	0.63

8. Gradient boosting classifier

Test accuracy	0.8172
ROC-AUC score	0.80

9. XG Boost classifier

Test accuracy	0.8180
ROC-AUC score	0.80

10. Cat Boost classifier

Test accuracy	0.8181
ROC-AUC score	0.80

Conclusion

Through experimental results, it was observed that all boosting algorithms—the CatBoost algorithm, the Gradient Boosting algorithm, and the XGBoost algorithm—provided the best ROC-AUC score of 0.80.

However, the CatBoost model achieved a slightly higher test accuracy of 82.15%.

It is recommended that public health officials find a way to make the vaccine accessible to people regardless of health insurance status.

Additionally, because opinion on H1N1 vaccine effectiveness and H1N1 risk to health are highly influential in determining vaccination status, the health authorities should make educational outreach a priority.