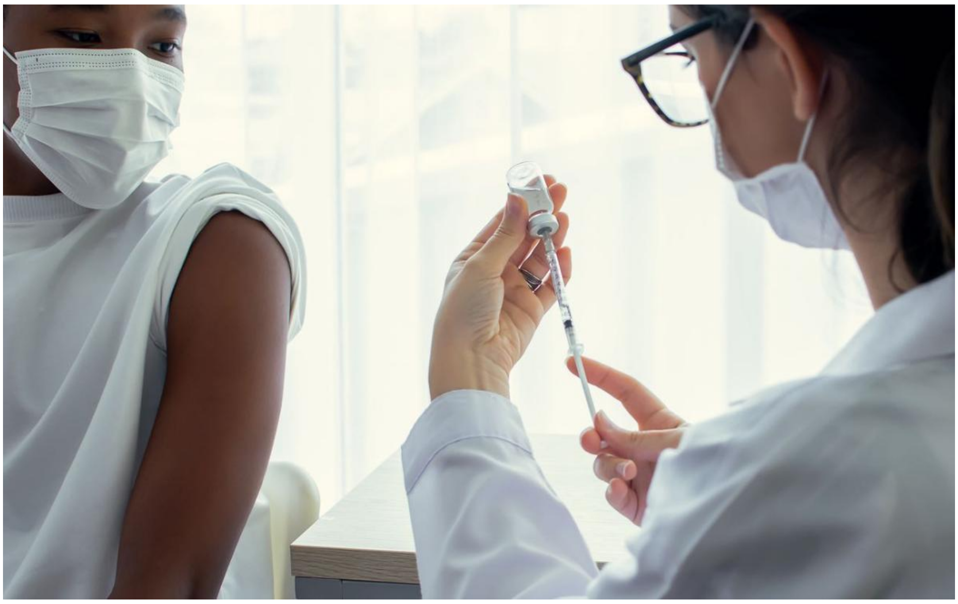
Report Science & Technology

PREDICTING H1N1 VACCINE UPTAKE

Phase Three Project On Predictive Modelling

Jennifer Njeri

23 October 2023



INTRODUCTION

The data for this project is sourced from the National 2009 H1N1 Flu Survey conducted in the United States following the Influenza outbreak of 2009. The datasets can be found on the Driven Data Website.

The dataset primarily consists of categorical variables with binary and numerical values. Additionally, certain columns contain coded data.

Project Objectives

1. Ideal Predictive Model: Develop a robust predictive model capable of estimating the likelihood of H1N1 vaccine uptake for individuals.
2. Feature Importance: Identify and prioritize the key features that contribute to the decision-making process regarding H1N1 vaccination.
3. Recommendations: Provide actionable insights to health authorities and policymakers to enhance targeted vaccination strategies.

MODEL SUCCESS CRITERIA

In the context of predicting vaccine intake, capturing as many true positive cases (individuals taking the vaccine) is crucial.

1. Recall, measuring the effectiveness of classification, is well-suited for this purpose. Identifying the characteristics of vaccine uptake informs targeted campaigns, allowing for efficient resource allocation and improved vaccination within specific demographics.
2. F1 Score, as a harmonic mean of precision and recall, ensures a balanced trade-off, sensitive to both false positives and false negatives. This aligns with the objectives of the prediction task.
3. Additionally, I'll use AUC-ROC to gauge the overall performance of the model.

Metrics:

1. AUC-ROC score of 85% and above
2. Balance Recall considering the target variable is heavily imbalanced. 50% and above.
3. 80% and above accuracy.
4. F1 Score of 50% and above

Columns description:

For all binary variables: 0 = No; 1 = Yes.

- h1n1_concern - Level of concern about the H1N1 flu. 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.
- h1n1_knowledge - Level of knowledge about H1N1 flu. 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.
- behavioral_antiviral_meds - Has taken antiviral medications. (binary)
- behavioral_avoidance - Has avoided close contact with others with flu-like symptoms. (binary)
- behavioral_face_mask - Has bought a face mask. (binary)
- behavioral_wash_hands - Has frequently washed hands or used hand sanitizer. (binary)
- behavioral_large_gatherings - Has reduced time at large gatherings. (binary)
- behavioral_outside_home - Has reduced contact with people outside of own household. (binary)
- behavioral_touch_face - Has avoided touching eyes, nose, or mouth. (binary)
- doctor_recc_h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
- doctor_recc_seasonal - Seasonal flu vaccine was recommended by doctor. (binary)
- chronic_med_condition - Has any of the following chronic medical conditions: asthma or an other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines
- taken for a chronic illness. (binary)
- child_under_6_months - Has regular close contact with a child under the age of six months. (binary)
- health_worker - Is a healthcare worker. (binary)
- health_insurance - Has health insurance. (binary)
- opinion_h1n1_vacc_effective - Respondent's opinion about H1N1 vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.

- opinion_h1n1_risk - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- opinion_h1n1_sick_from_vacc - Respondent's worry of getting sick from taking H1N1 vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
- opinion_seas_vacc_effective - Respondent's opinion about seasonal flu vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
- opinion_seas_risk - Respondent's opinion about risk of getting sick with seasonal flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- opinion_seas_sick_from_vacc - Respondent's worry of getting sick from taking seasonal flu vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
- age_group - Age group of respondent.
- education - Self-reported education level.
- race - Race of respondent.
- sex - Sex of respondent.
- income_poverty - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- marital_status - Marital status of respondent.
- rent_or_own - Housing situation of respondent.
- employment_status - Employment status of respondent.
- hhs_geo_region - Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
- census_msa - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
- household_adults - Number of other adults in household, top-coded to 3.
- household_children - Number of children in household, top-coded to 3.

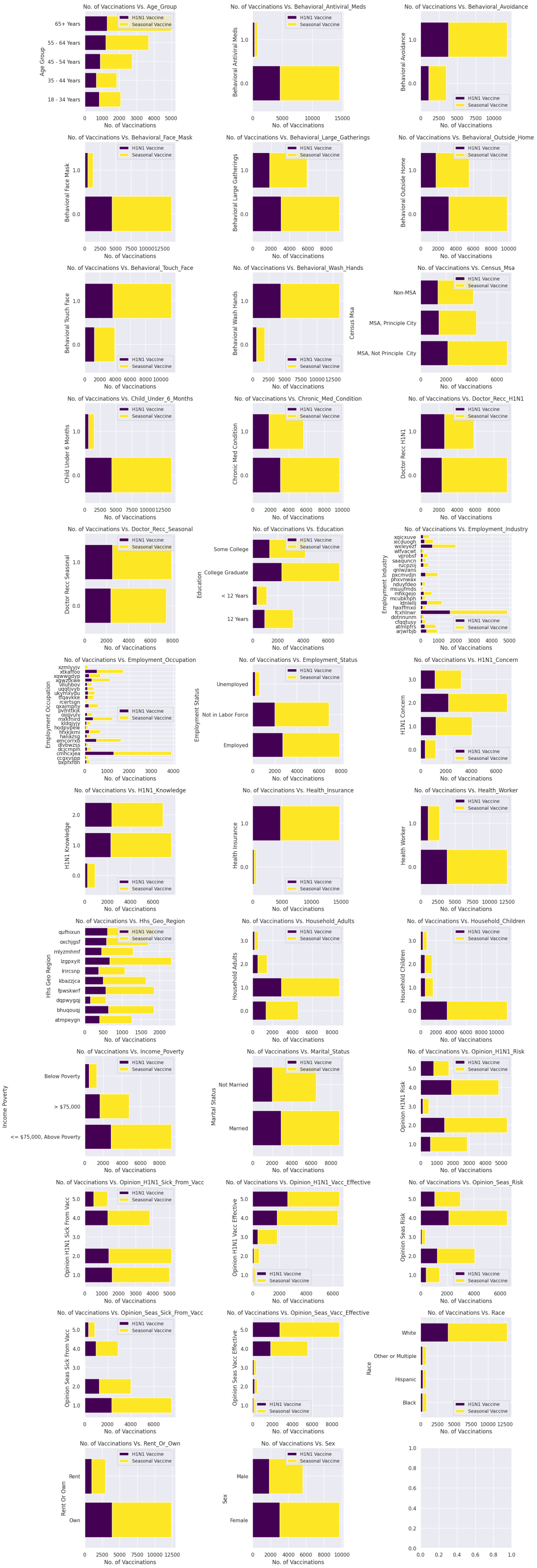
- employment_industry - Type of industry respondent is employed in. Values are represented as short random character strings.
- employment_occupation - Type of occupation of respondent. Values are represented as short random character strings.

DATAFRAME DESCRIPTION

```
Int64Index: 26707 entries, 0 to 26706
Data columns (total 38 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   respondent_id                26707 non-null  int64
 1   h1n1_vaccine                 26707 non-null  int64
 2   seasonal_vaccine             26707 non-null  int64
 3   h1n1_concern                 26615 non-null  float64
 4   h1n1_knowledge               26591 non-null  float64
 5   behavioral_antiviral_meds    26636 non-null  float64
 6   behavioral_avoidance         26499 non-null  float64
 7   behavioral_face_mask         26688 non-null  float64
 8   behavioral_wash_hands        26665 non-null  float64
 9   behavioral_large_gatherings  26620 non-null  float64
 10  behavioral_outside_home      26625 non-null  float64
 11  behavioral_touch_face        26579 non-null  float64
 12  doctor_recc_h1n1             24547 non-null  float64
 13  doctor_recc_seasonal         24547 non-null  float64
 14  chronic_med_condition        25736 non-null  float64
 15  child_under_6_months         25887 non-null  float64
 16  health_worker                25903 non-null  float64
 17  health_insurance             14433 non-null  float64
 18  opinion_h1n1_vacc_effective  26316 non-null  float64
 19  opinion_h1n1_risk            26319 non-null  float64
 20  opinion_h1n1_sick_from_vacc  26312 non-null  float64
 21  opinion_seas_vacc_effective  26245 non-null  float64
 22  opinion_seas_risk            26193 non-null  float64
 23  opinion_seas_sick_from_vacc  26170 non-null  float64
 24  age_group                    26707 non-null  object
 25  education                    25300 non-null  object
 26  race                         26707 non-null  object
```

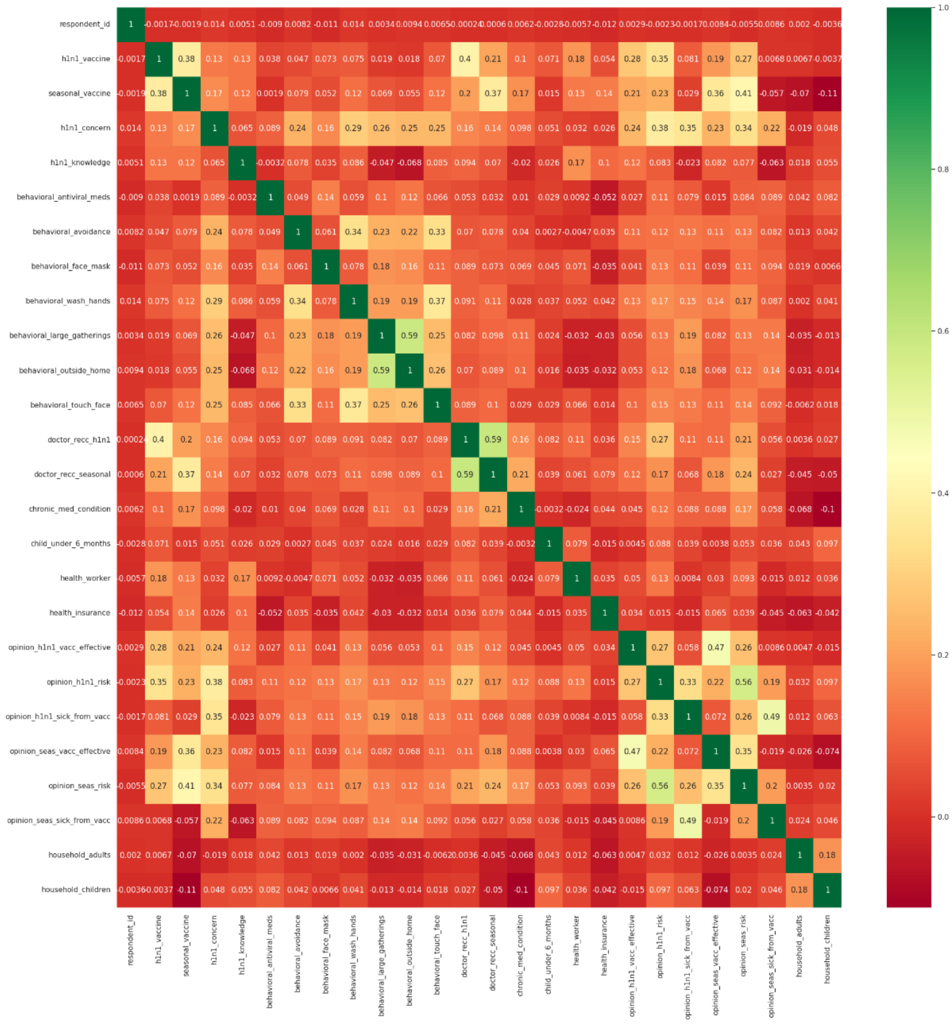
```
 27  sex                        26707 non-null  object
 28  income_poverty             22284 non-null  object
 29  marital_status             25299 non-null  object
 30  rent_or_own                24665 non-null  object
 31  employment_status          25244 non-null  object
 32  hhs_geo_region             26707 non-null  object
 33  census_msa                 26707 non-null  object
 34  household_adults           26458 non-null  float64
 35  household_children         26458 non-null  float64
 36  employment_industry        13377 non-null  object
 37  employment_occupation      13237 non-null  object
dtypes: float64(23), int64(3), object(12)
memory usage: 7.9+ MB
```

DISTRIBUTION OF H1N1 VACCINE COMPARED TO SEASONAL VACCINE



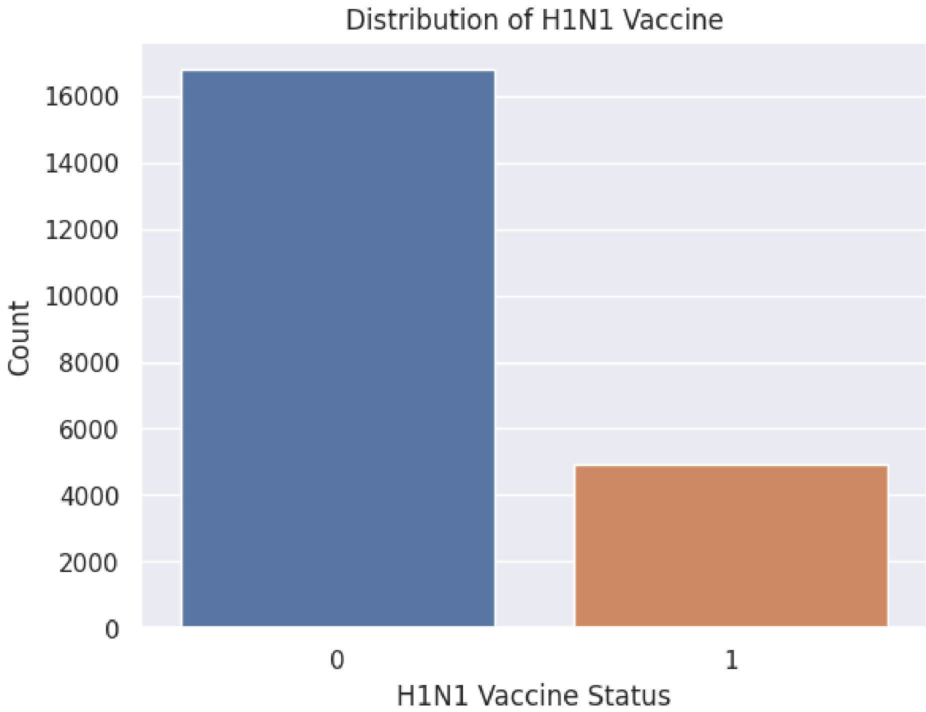
More individuals are opting for Seasonal Flue Vaccine over the H1N1 Vaccine.

CORRELATION BETWEEN VARIABLES



The correlations observed against the H1N1 Vaccine are within moderate levels, suggesting potential compatibility for regression modeling without encountering multicollinearity issues. Lasso and Ridge Regularization shall solve for any multicollinearity.

CLASS IMBALANCE

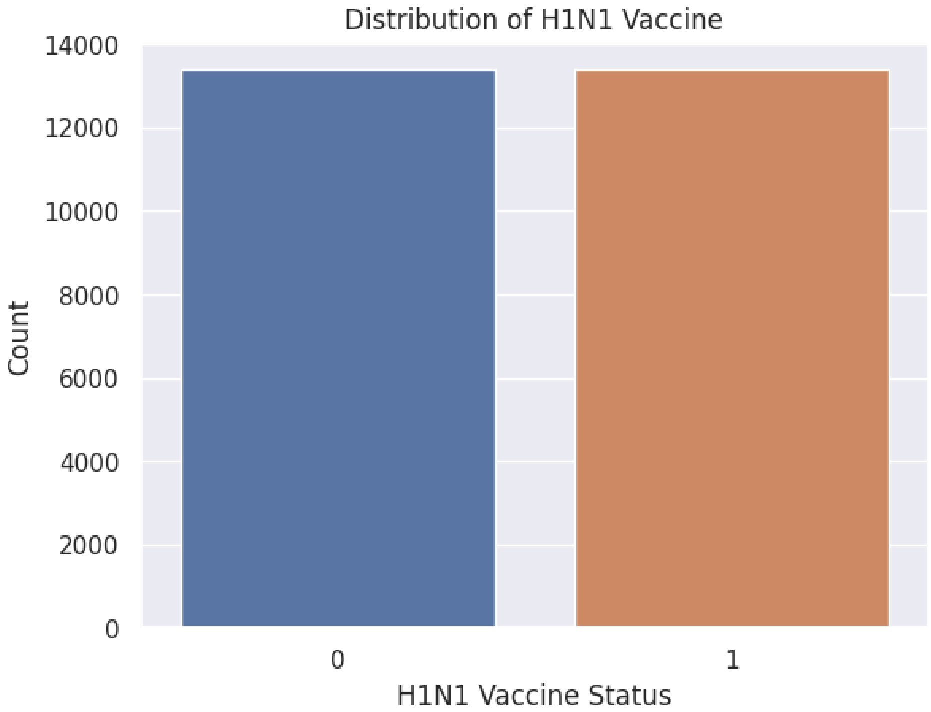


There is class imbalance in our target variable

Solve for class imbalance

Plain text

```
# Apply SMOTE to balance the classes in the
training set
smote = SMOTE(random_state=42)
X_train_smote, y_train_smote =
smote.fit_resample(X_train, y_train)
```



Class Imbalance Solved

SCALE THE DATA

Plain text

```
# Use Standard Scalar to Scale the data
scaler = StandardScaler()

# Fit the scaler on the SMOTE data
X_train_scaled =
scaler.fit_transform(X_train_smote)
X_test_scaled = scaler.transform(X_test)
```

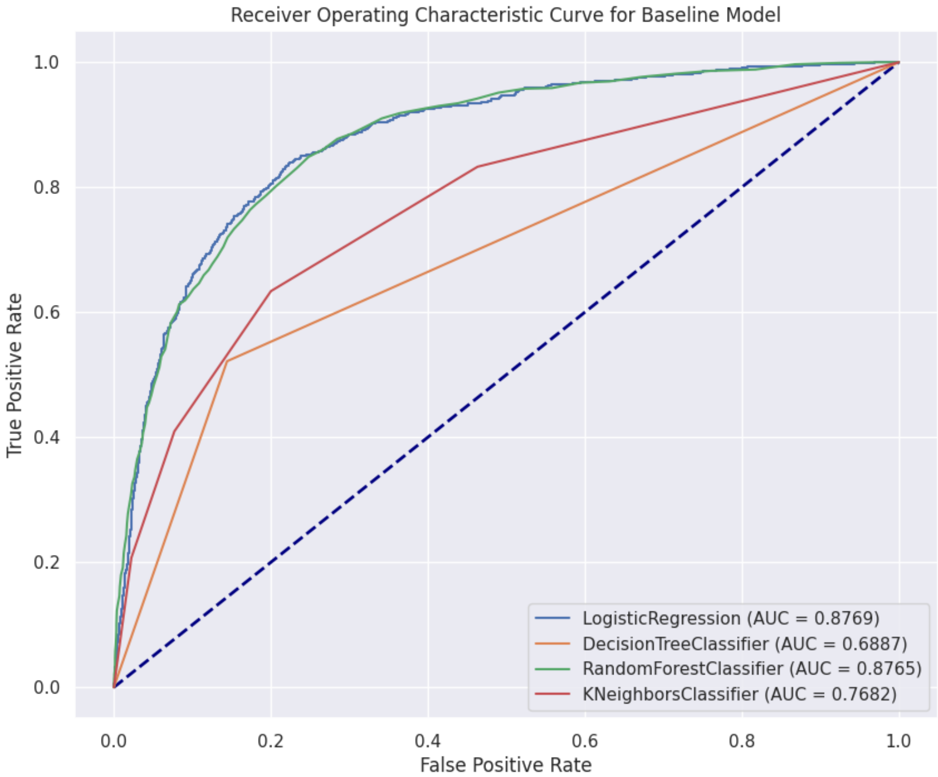
BASELINE MODEL

AUC-ROC SCORES

LogisticRegression - AUC-ROC: 0.8769
DecisionTreeClassifier - AUC-ROC: 0.6887
RandomForestClassifier - AUC-ROC: 0.8765
KNeighborsClassifier - AUC-ROC: 0.7682

Classification scores

LogisticRegression - Accuracy: 0.8535, Recall: 0.5672, F1 Score: 0.6276, AUC-ROC: 0.8769
DecisionTreeClassifier - Accuracy: 0.7830, Recall: 0.5217, F1 Score: 0.5114, AUC-ROC: 0.6887
RandomForestClassifier - Accuracy: 0.8478, Recall: 0.4825, F1 Score: 0.5798, AUC-ROC: 0.8765
KNeighborsClassifier - Accuracy: 0.8111, Recall: 0.4095, F1 Score: 0.4856, AUC-ROC: 0.7682



Visualize the auc-roc curves of the baseline models.

Considering our objectives, the models which prioritize Recall, F1 Score, and AUC-ROC are:

1. Logistic Regression has the highest Recall, F1 Score, and AUC-ROC among all models, making it a strong candidate.
2. Random Forest Classifier has a good balance of Accuracy, Recall, and F1 Score. The AUC-ROC is also high.

Modelling with balanced (SMOTE) and Scaled (StandardScaler) data

We will use Logistic Regression and Random Forest Classifiers.

Classification Report

Plain text

```
Logistic Regression Classification Report:
              precision    recall  f1-score
support

           0       0.87      0.95      0.91
3397
           1       0.72      0.51      0.60
945

    accuracy                           0.85
4342
   macro avg       0.80      0.73      0.75
4342
weighted avg       0.84      0.85      0.84
4342

Logistic Regression AUC-ROC:
0.8747618891863814

Random Forest Classification Report:
              precision    recall  f1-score
support

           0       0.88      0.93      0.90
3397
           1       0.68      0.55      0.61
945

    accuracy                           0.85
4342
   macro avg       0.78      0.74      0.76
4342
weighted avg       0.84      0.85      0.84
4342

Random Forest AUC-ROC: 0.8723036043318646
```

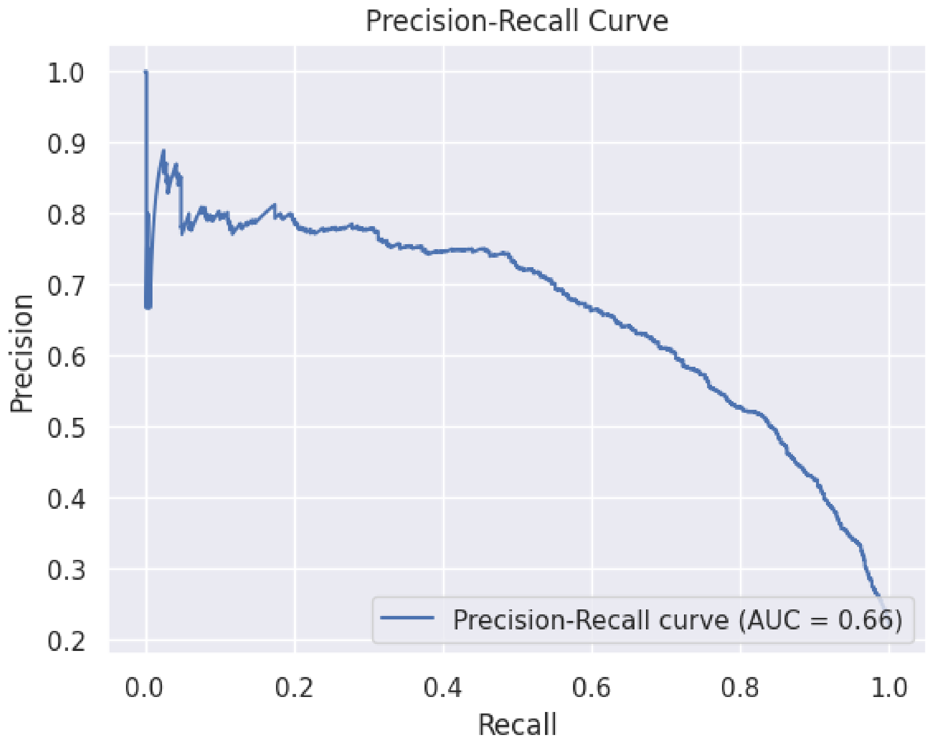
Overall Performance

Logistic Regression Metrics:
Accuracy: 0.8505
Recall: 0.5079
F1 Score: 0.5966
Random Forest Metrics:
Accuracy: 0.8452
Recall: 0.5524
F1 Score: 0.6084

The Logistic Regression performs slightly lower than the Random Forest model in terms of recall. Let us boost the recall scores in the Logistic regression model.
While the default recall threshold is typically set at 0.5 for binary classification to achieve a balance, the goal here is to optimize TPR performance. Thus, we will fine-tune the threshold, taking into consideration the importance of precision as well.

Plain text

```
# Logistic Regression
# Explore different thresholds
thresholds = [0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
0.8]
for threshold in thresholds:
    y_pred_adjusted = (y_pred_proba_logreg
> threshold).astype(int)
    print(f"Threshold: {threshold}")
    print(classification_report(y_test,
y_pred_adjusted))

# Optimal threshold based on the objectives
optimal_threshold = 0.2

# Apply the optimal threshold to get final
predictions
y_pred_final = (y_pred_proba_logreg >
optimal_threshold).astype(int)

# Evaluate the model performance with the
optimal threshold
print("Final Classification Report:")
print(classification_report(y_test,
y_pred_final))

# Visualize precision-recall curve
precision, recall, thresholds =
precision_recall_curve(y_test,
y_pred_proba_logreg)
area_under_curve = auc(recall, precision)

plt.plot(recall, precision,
label=f'Precision-Recall curve (AUC =
{area_under_curve:.2f})')
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title('Precision-Recall Curve')
plt.legend(loc="lower right")
plt.show()
```



As Recall threshold increases, the precision decreases. We will use a threshold of 0.2 to strick a balance between precision and recall.

Hyperparameter Tuning on Logistic Regression and XGBOOST

We will use GridSearchCV to find the best hyperparameters for our Logistic Regression model, Random Forest and XGBoost model and assess their performance on the test set.

Classification Report

Plain text

```
Classification Report - Logistic
Regression:
               precision    recall  f1-
score   support

           0       0.98      0.14      0.24
3397
           1       0.24      0.99      0.39
945

    accuracy                           0.32
4342
   macro avg       0.61      0.56      0.31
4342
weighted avg       0.82      0.32      0.27
4342

Classification Report - Random Forest:
               precision    recall  f1-
score   support

           0       0.00      0.00      0.00
3397
           1       0.22      1.00      0.36
945

    accuracy                           0.22
4342
   macro avg       0.11      0.50      0.18
4342
weighted avg       0.05      0.22      0.08
4342

Classification Report - XGBoost:
               precision    recall  f1-
score   support

           0       0.88      0.94      0.91
3397
           1       0.70      0.53      0.61
945
```

```
    accuracy                           0.85
4342
   macro avg       0.79      0.74      0.76
4342
weighted avg       0.84      0.85      0.84
4342
```

Overall Performance

Best Logistic Regression Model Metrics:
Accuracy: 0.3217
Recall: 0.9894
F1 Score: 0.3884
AUC-ROC: 0.8488
Best XGBoost Model Metrics:
Accuracy: 0.8489
Recall: 0.5344
F1 Score: 0.6062
AUC-ROC: 0.8796
Best Random Forest Model Metrics:
Accuracy: 0.2176
Recall: 1.0000
F1 Score: 0.3575
AUC-ROC: 0.4570

Observation

Logistic Regression:

High recall for class 1 (0.99), indicating it correctly identifies positive instances. Low precision for class 1 (0.24), suggesting a high number of false positives. Overall low accuracy (0.32).

Random Forest:

Perfect recall for class 1 (1.00), meaning it correctly identifies all positive instances. Low precision for class 1 (0.22), indicating a high number of false positives. Extremely low accuracy (0.22).

XGBoost:

Balanced recall (0.53) and precision (0.70) for class 1. Higher overall accuracy (0.85) compared to the other models.

Reasoning:

While Random Forest has perfect recall for class 1, its precision is very low, leading to a high number of false positives and low accuracy. XGBoost strikes a balance between recall, precision, and accuracy. It performs well across all metrics.

Objective One: Ideal Predictive Model

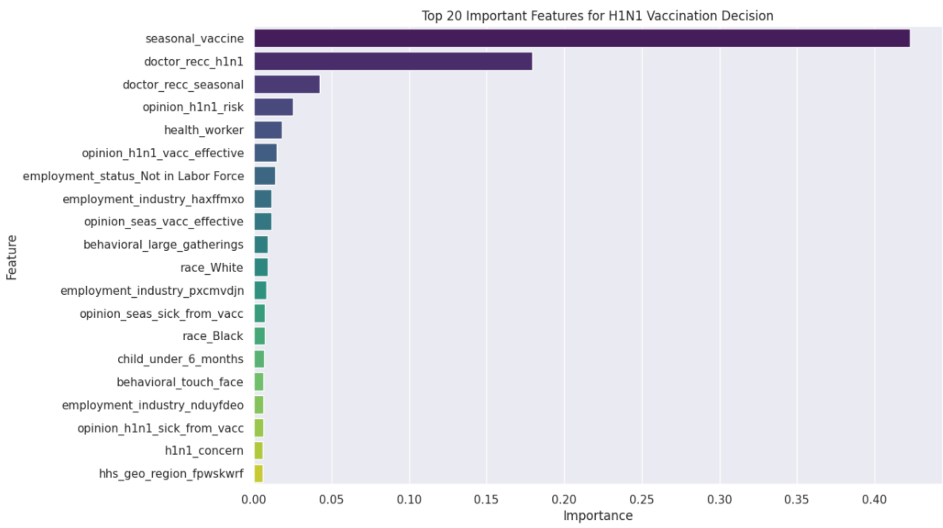
Develop a robust predictive model capable of estimating the likelihood of H1N1 vaccine uptake for individuals.
XGBoost seems to be the best-performing model among the three, considering a balance between precision, recall, and accuracy. It offers a better trade-off between correctly identifying positive instances and minimizing false positives.

Objective Two: Identify and prioritize the key features that contribute to the decision-making process regarding H1N1 vaccination.

Plain text

```
# Extract feature importance scores
feature_importance =
best_xgb_model.feature_importances_

# Associate feature names with importance
scores
feature_names = X_train.columns
feature_importance_dict =
dict(zip(feature_names,
feature_importance))

# Sort features by importance
sorted_features =
sorted(feature_importance_dict.items(),
key=lambda x: x[1], reverse=True)

# Display the top 20 features
top_features = sorted_features[:20]
top_features_df =
pd.DataFrame(top_features, columns=
['Feature', 'Importance'])
top_features_df
```

Top 20 important features:



These are the factors that influence H1N1 Vaccine Uptake.

Objective three: Recommendations: Provide actionable insights to health authorities and policymakers to enhance targeted vaccination strategies.

1. Encourage Seasonal Vaccine Uptake: Given that seasonal_vaccine is the most important feature, public health campaigns should emphasize and promote the importance of receiving the seasonal flu vaccine.
2. Promote Doctor Recommendations: As doctor_recc_h1n1 and doctor_recc_seasonal are significant, efforts should be made to enhance communication between healthcare professionals and the public. Encourage doctors to recommend both H1N1 and seasonal flu vaccines during patient visits.
3. Address Perceived Risks: Since opinion_h1n1_risk and opinion_seas_sick_from_vacc are influential, public health messaging should address and clarify any misconceptions or concerns regarding the perceived risks associated with H1N1 and seasonal flu vaccinations.
4. Target Health Workers: The importance of health_worker as a feature suggests that targeting healthcare workers for vaccination campaigns and ensuring their high vaccination rates could positively influence the general public.
5. Effective Communication Strategies: Recognizing the impact of opinions on vaccine effectiveness (opinion_h1n1_vacc_effective and opinion_seas_vacc_effective), public health campaigns should employ clear and compelling communication strategies to convey the effectiveness of both H1N1 and seasonal flu vaccines.
6. Employment Status Considerations: The feature employment_status_Not in Labor Force is significant. Tailoring vaccination campaigns to different employment statuses and addressing barriers specific to those not in the labor force could improve overall vaccine uptake.
7. Diversity and Racial Considerations: The features race_White and race_Black suggest considering diversity and tailoring campaigns to specific racial or ethnic groups to ensure inclusivity and effectiveness.

8. Behavioral Interventions: Focusing on behavioral aspects, such as behavioral_large_gatherings and behavioral_touch_face, indicates the importance of interventions promoting preventive behaviors in high-risk situations.
9. Child Vaccination Considerations: Given that child_under_6_months is a significant feature, campaigns should address concerns and provide information about the safety and importance of vaccinating children under six months.
10. Income and Economic Considerations: Acknowledging the importance of income_poverty, addressing economic barriers and offering accessibility to free or low-cost vaccination services can contribute to increased vaccine uptake.

About the author

Jennifer Njeri