# Vaccination Willingness Classification

Hyewon Jeong
Metis

# Motivation

- Vaccination is the most effective way to prevent infectious diseases
- Despite evidence of the effectiveness and safety of vaccinations, a number of people in the U.S. are reluctant to receive vaccinations.

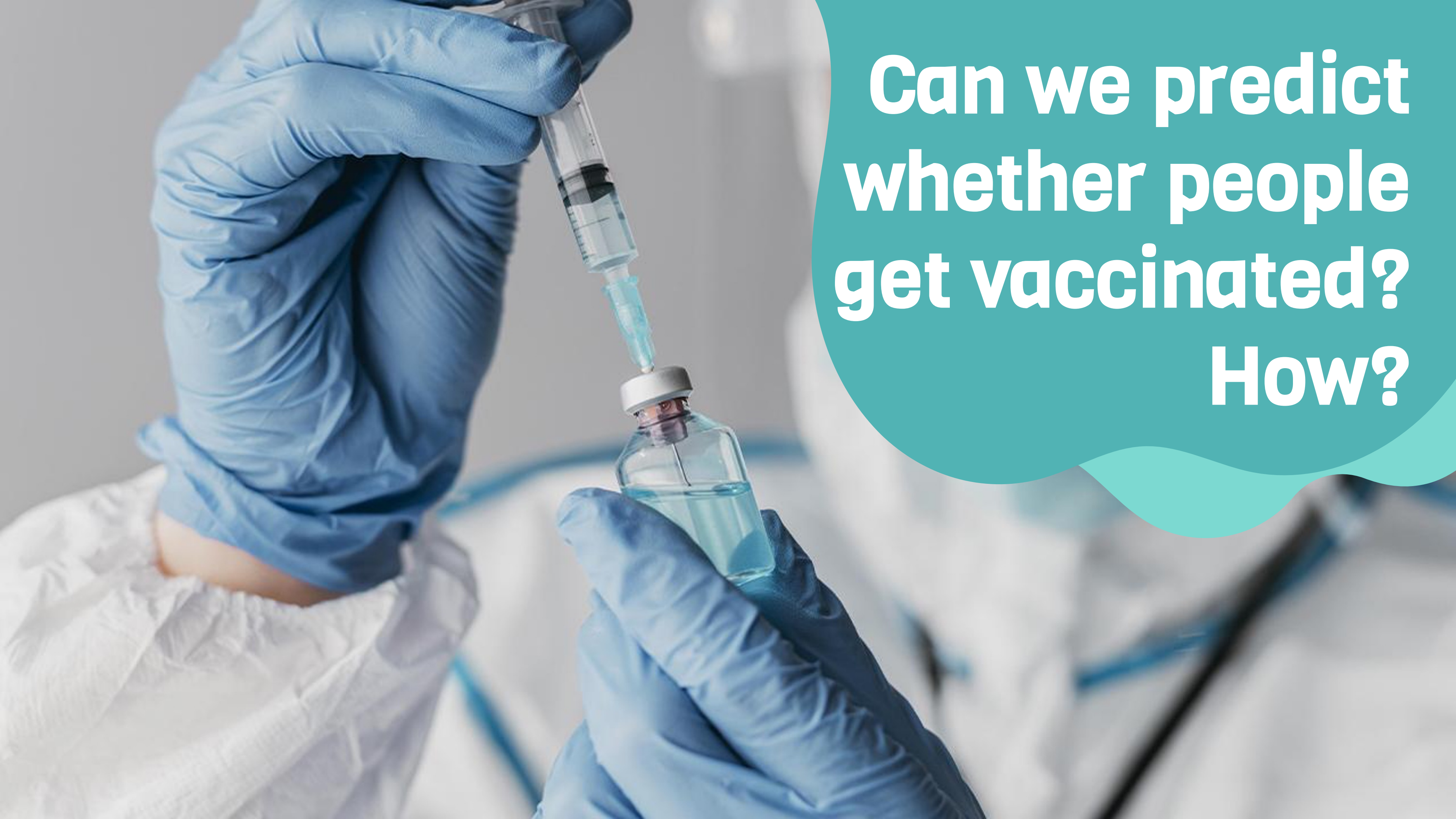
# Vaccine hesitancy

- Vaccine hesitancy gained worldwide attention as the COVID-19 pandemic swept the globe

- Many citizen in the United States were unwilling to take vaccinations against COVID-19 due to the following reasons:
  - Concerns about safety
  - Side effects
  - General mistrust of the government

**27 %**

Of adults in the U.S. would definitely not get a COVID-19 vaccination



# Can we predict whether people get vaccinated? How?

# About Data

## The National 2009 H1N1 Flu Survey

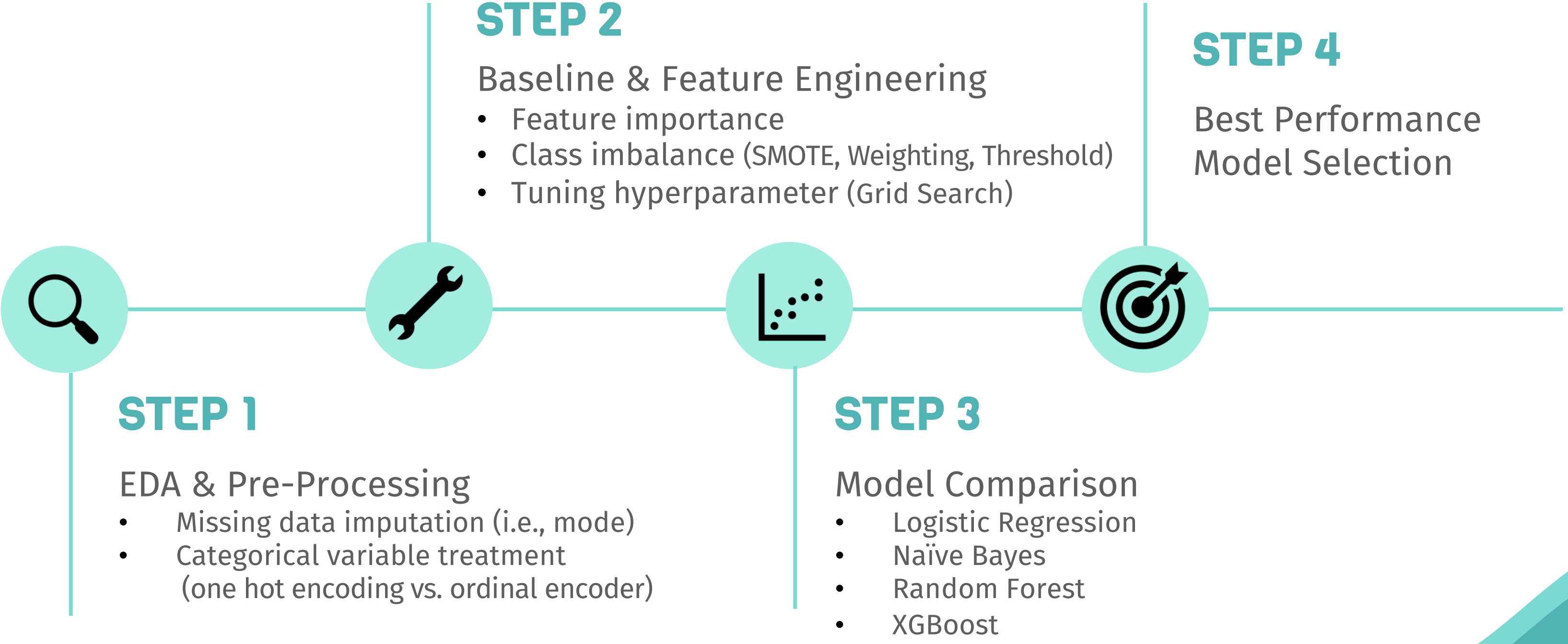
A phone survey asked respondents whether they had received the H1N1 vaccine (swine flu), including followings:

- social, economic, and demographic backgrounds
- opinions on risks of illness and vaccine effectiveness
- behaviors towards mitigating transmission

## Details

- Target: whether vaccinated for H1N1
    - 22 % of target variables - positive
- 35 Categorical Features

# Methodology



## STEP 1

EDA & Pre-Processing

- Missing data imputation (i.e., mode)
- Categorical variable treatment (one hot encoding vs. ordinal encoder)

## STEP 2

Baseline & Feature Engineering

- Feature importance
- Class imbalance (SMOTE, Weighting, Threshold)
- Tuning hyperparameter (Grid Search)

## STEP 3

Model Comparison

- Logistic Regression
- Naïve Bayes
- Random Forest
- XGBoost

## STEP 4

Best Performance Model Selection

# Baseline

## Logistic Regression

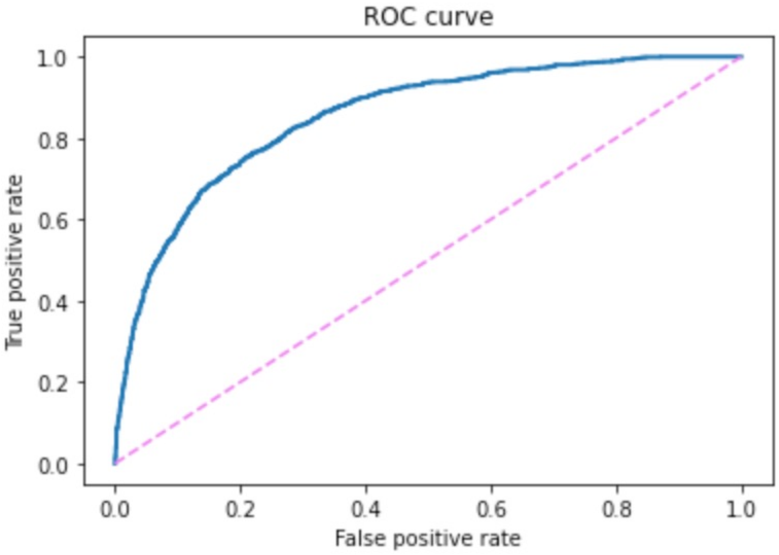
Score on Training : 0.835

Score on Validation: 0.848

Good!

Recall Score: 0.442

ROC AUC Score: 0.854



# Notable Feature Engineering:

## Class imbalance + Grid Search CV

### Logistic Regression

Recall Score

0.461 → 0.730

### Naïve Bayes

Recall Score

0.573 → 0.687

### Random Forest

Recall Score

0.584 → 0.742

# Model Comparison

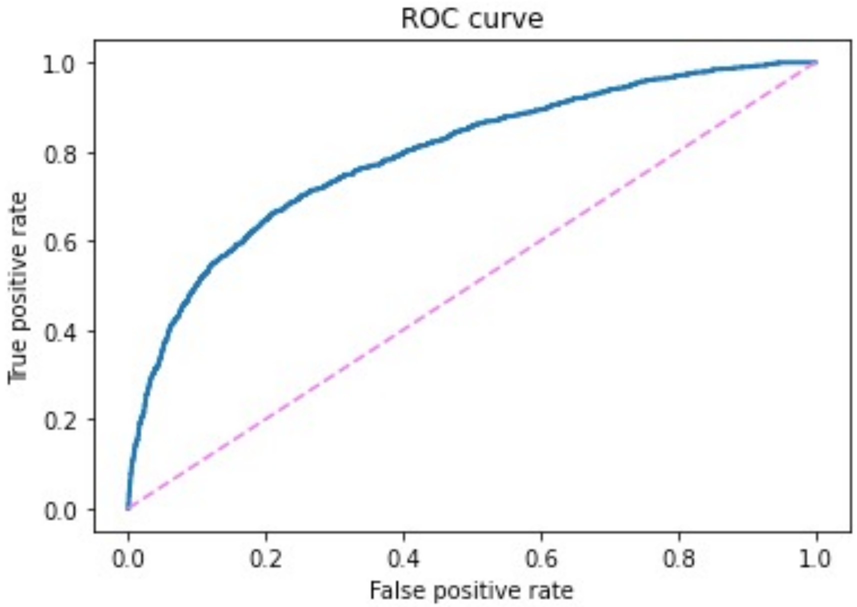
(hyperparameters in all models except for Baseline were tuned with Grid Search CV)

| Scores | Train | Validation | Recall |
|---|---|---|---|
| Baseline | 0.835 | 0.848 | 0.442 |
| Logistic Regression (with selected features) | 0.829 | 0.843 | 0.696 |
| Logistic Regression (with all features) | 0.835 | 0.843 | 0.730 |
| Naïve Bayes | 0.791 | 0.801 | 0.687 |
| Random Forest | 0.740 | 0.745 | 0.742 |
| XGBoost | 0.850 | 0.844 | 0.470 |

# Best Performance Model

## Random Forest

The Score on Test Data: 0.751
The Score on Recall : 0.701

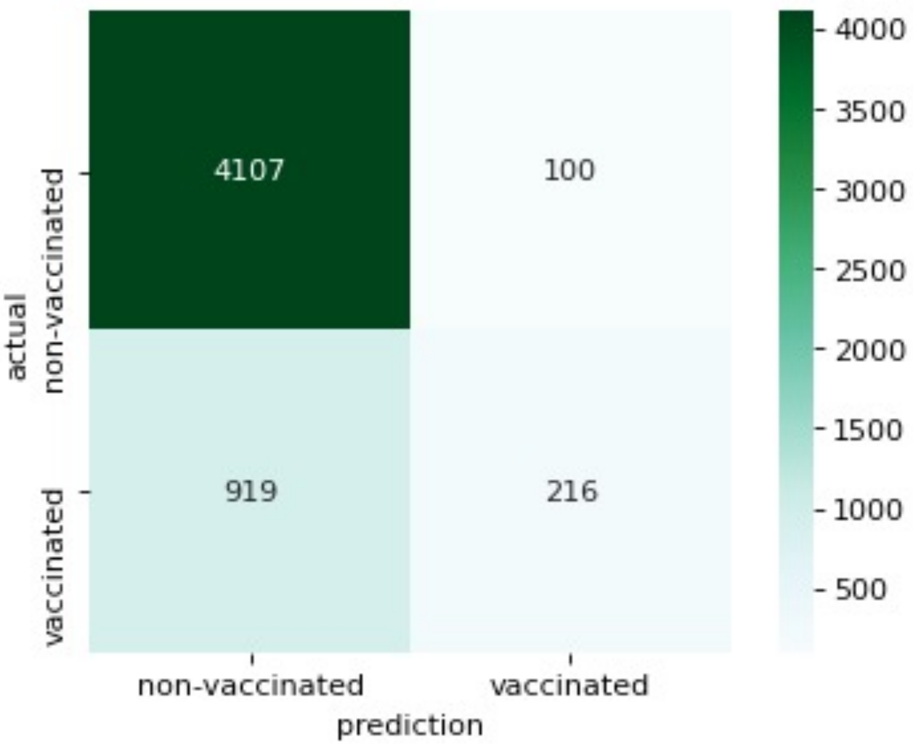


**ROC AUC score: 0.792**

# Best Performance Model

## Random Forest

The Score on Test Data: 0.751
The Score on Recall : 0.701



# Future Studies

## Advanced Feature Engineering 01

- Feature selection
- Further aggregating or grouping categories
  - Employment occupation and industry
  - geolocation

## Other Ensembling Methods 02

# Thank you.