PATRICK UWIGIZE
UDOBANG JOSHUA JACOB

FLU SHOT LEARNING: PREDICT H1N1 AND SEASONAL FLU VACCINES

TEAMS: 28

Skoltech

01

PROBLEM DESCRIPTION

The COVID-19 pandemic has brought vaccines and immunization to the forefront of public health strategies. Vaccination plays a crucial role in preventing the spread of viral diseases like COVID-19 by achieving herd immunity. Examining the 2009 H1N1 influenza pandemic, a highly contagious respiratory illness, provides insights into vaccination rates and their correlation with personal traits. The National 2009 H1N1 Flu Survey data collected during the vaccination campaign helps understand the influence of demographics and individual beliefs on immunization trends. By analyzing this data, future public health initiatives can be directed more effectively. A model has been developed to understand the psychological factors affecting vaccine uptake, make predictions, and establish guidelines for vaccination. This model enhances preparedness for future pandemics, improves vaccination quality, reduces information gathering costs, and aids in mitigating outbreaks and saving lives.

02

LOADING DATA

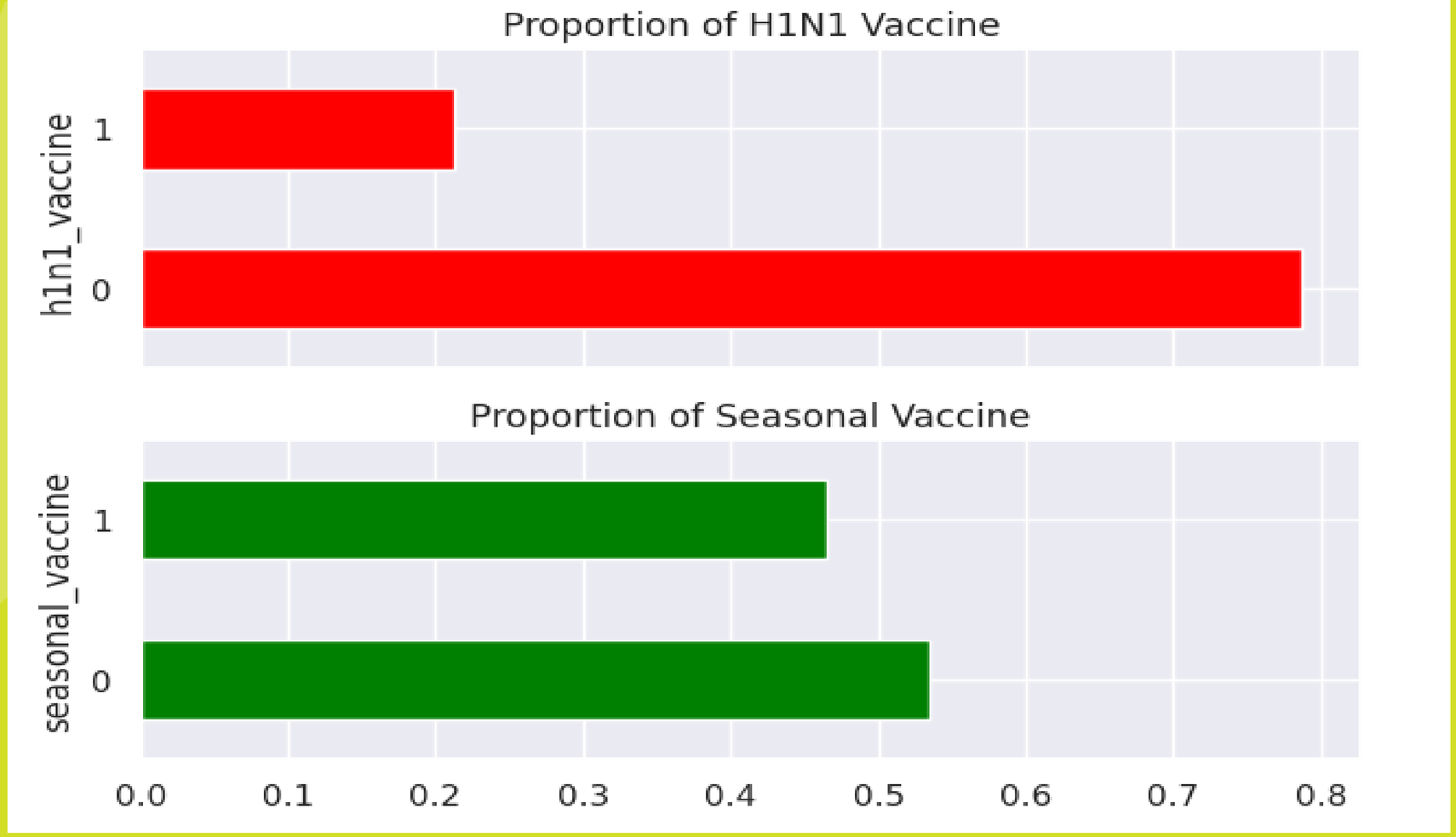
- Load the data for analysis.
- Observe the top 10 feature observations to gain insights into the dataset.
- Observe the last 10 feature observations to gain insights into the dataset.
- Print all the columns/features names for reference.
- Print data size to understand the size of the dataset.

03

DATA EXPLORATION

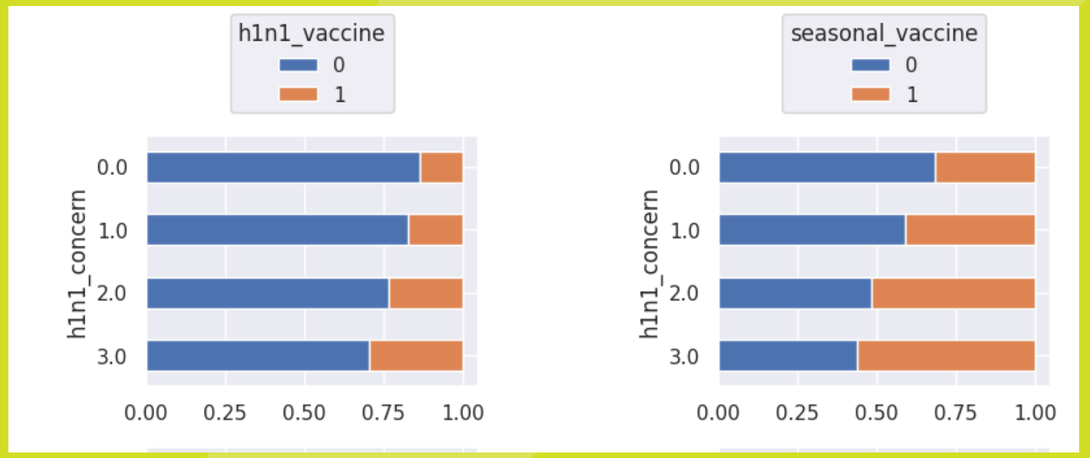
1. Data Cleaning: This involved removing or correcting any errors, missing values, or inconsistencies in the data (imputing missing values, removing duplicates, and handling outliers)
2. Data Transformation: into a format that can be used for analysis. Converting raw data.
3. Feature Selection: This involved selecting a subset of relevant features that are most useful for the analysis.
4. Data Reduction: This involves reducing the size of the dataset to improve computational efficiency and reduce overfitting.
5. Data Splitting: This involved splitting the dataset into training, validation, and testing sets to evaluate the performance of the model on unseen data.

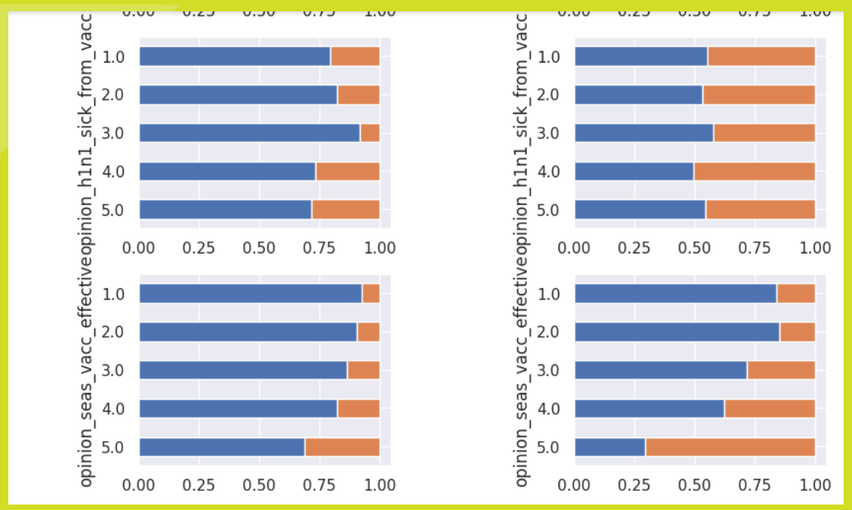
04

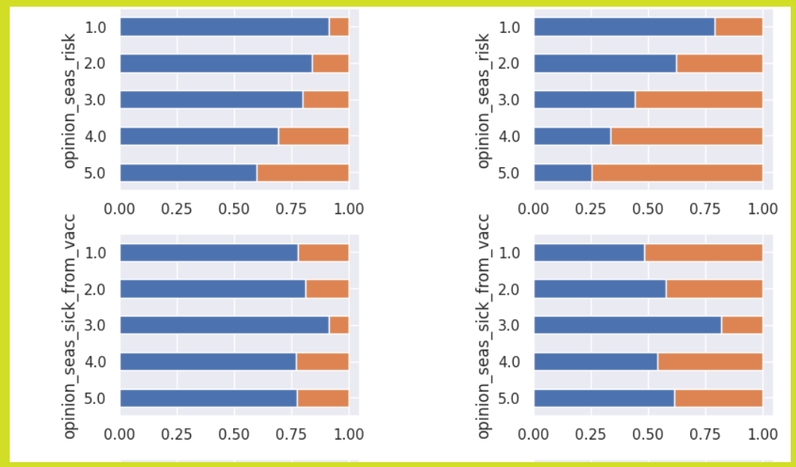


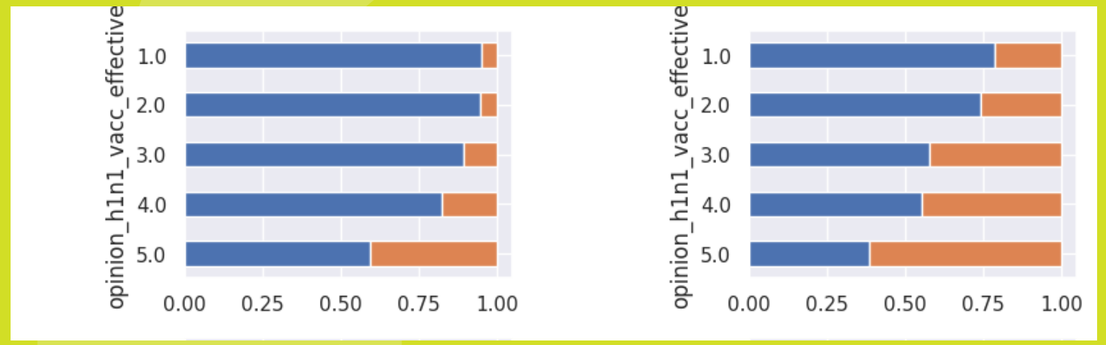
About 20% of persons appear to have got the H1N1 flu vaccine,compared to roughly 50% who had the seasonal flu shot.We characterize the seasonal flu vaccine goal as having balanced classes and the H1N1 flu vaccine target as having somewhat imbalanced classes in terms of class balance.

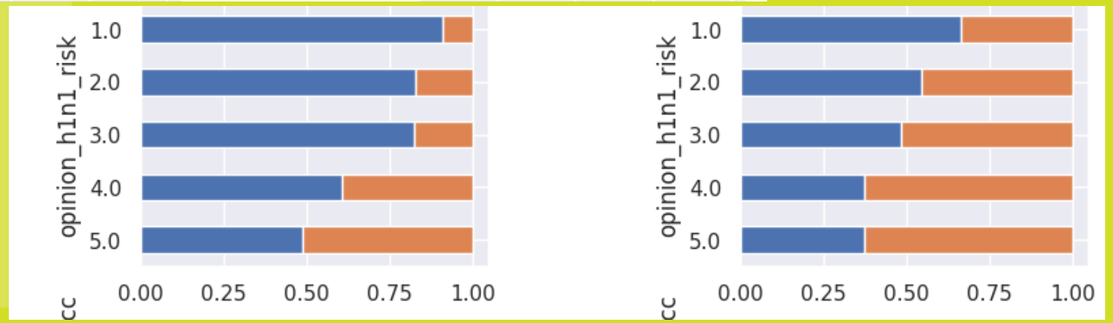
05


h1n1_vaccine
0
1
seasonal_vaccine
0
1
h1n1_concern
0.0
1.0
2.0
3.0
0.00
0.25
0.50
0.75
1.00


opinion_h1n1_sick_from_vacc
opinion_seas_vacc_effective
1.0
2.0
3.0
4.0
5.0
0.00
0.25
0.50
0.75
1.00


opinion_seas_risk
opinion_seas_sick_from_vacc
1.0
2.0
3.0
4.0
5.0
0.00
0.25
0.50
0.75
1.00


opinion_h1n1_vacc_effective
1.0
2.0
3.0
4.0
5.0
0.00
0.25
0.50
0.75
1.00


opinion_h1n1_risk
1.0
2.0
3.0
4.0
5.0
0.00
0.25
0.50
0.75
1.00


sex
Female
Male
age_group
18 - 34 Years
35 - 44 Years
45 - 54 Years
55 - 64 Years
65+ Years
0.00
0.25
0.50
0.75
1.00

06

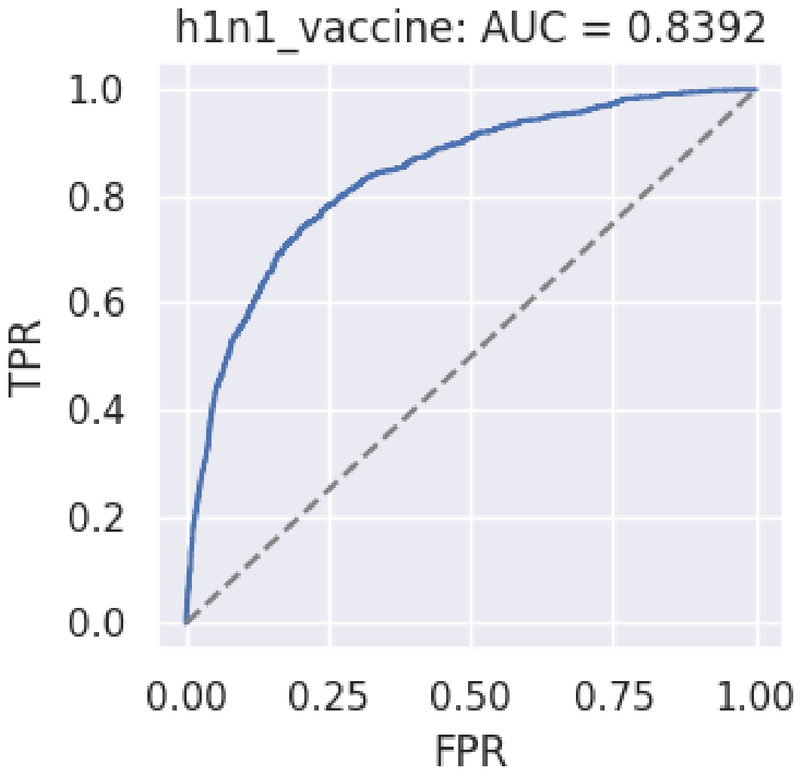
The knowledge and opinion questions appear to have a fairly strong signal for both of the target variables. The seasonal vaccine and the demographic characteristics are more strongly correlated, while the h1n1 vaccine is still far less so. Specifically, it is noteworthy to observe a robust association between age group and the seasonal vaccination, but not with the h1n1 vaccine. People seem to respond adequately to seasonal influenza, given that older individuals are more susceptible to consequences from the virus. However, it turns out that there is an intriguing correlation between age and H1N1 flu: older individuals were less likely to contract the virus despite having a higher risk of consequences! If this research tells us anything about causality, it appears that the risk variables ultimately found their way into the vaccination rates.

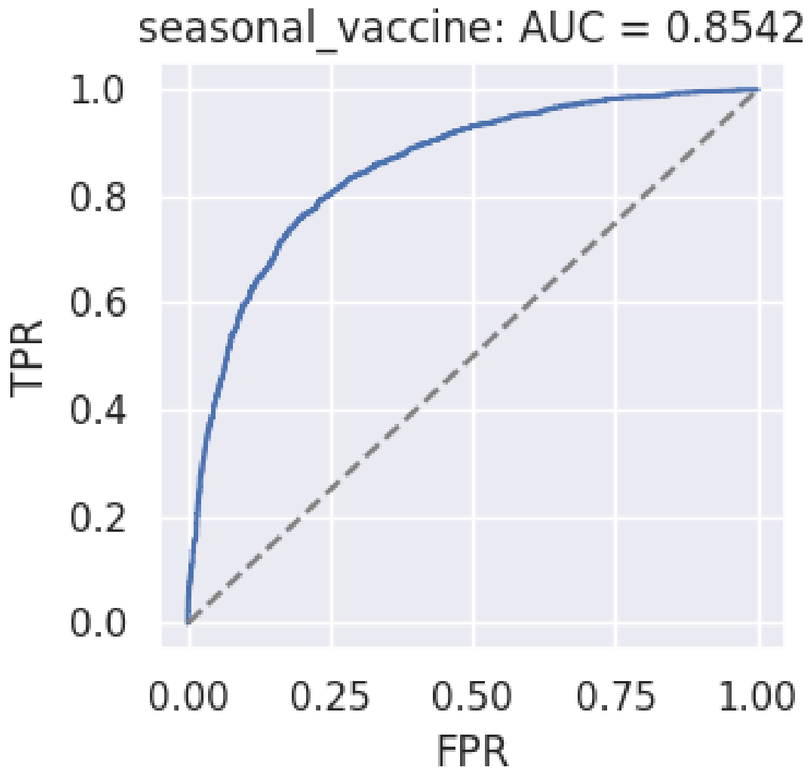
PREPROCESSING

- Drop target columns from training data
- Adding one-hot encoding columns and dropping initial categorical columns from the dataframe.

BUILDING MODELS

- Logistic Regression

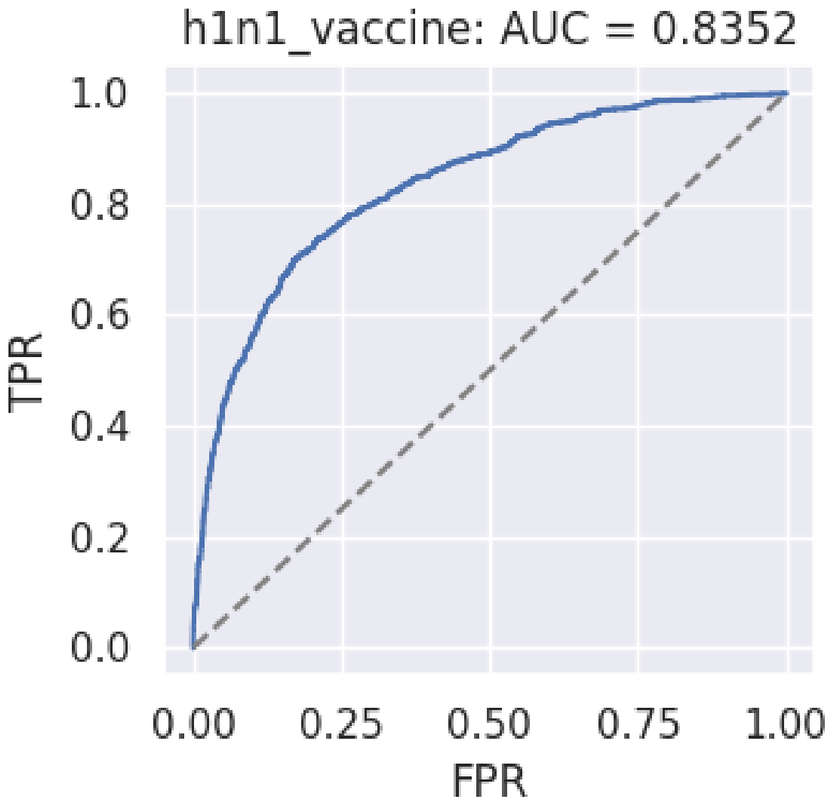


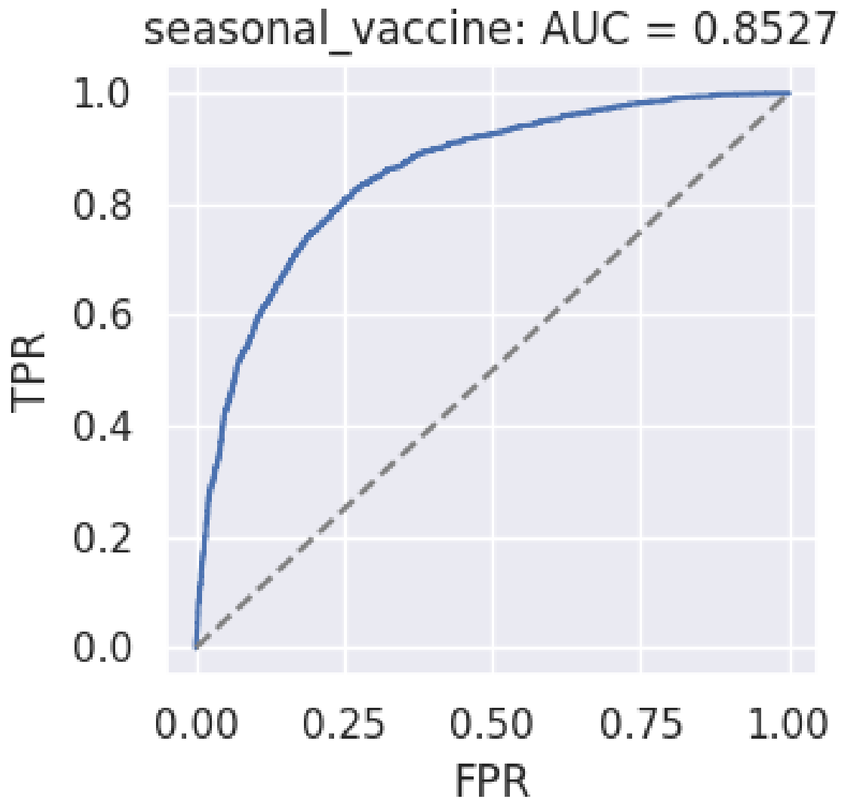


Auc_roc score - 0.847

BUILDING MODELS

- Random Forest





Auc_roc score - 0.844

CONCLUSION

Based on the prediction from both models, the logistic regression result seems to beat the Random Forest result thus, we go with the logistic regression model with auc_roc_score 0.847 which is a good score as it is close to 1 and shows a high correlation between the training features and the target variables which is really nice.
From the correlation matrix, we were able to fish out the most correlated. Parameters with the target labels and make reasonable predictions with them. Thus, we achieved our initial goal as intended.

Thx!

Skoltech