bnittalee / H1N1-Flu-Vaccines Public

Phase 3 project

0 stars 0 forks

Star Watch

Code Issues Pull requests Actions Projects Security Insights Se

main

bnittalee Update notebook ... 5 days ago 25

View code

README.md

Developing and Comparing Machine Learning Models for Predicting Flu Vaccine Likelihood

By Brittney Nitta-Lee

Business and Data Understanding

The National Center for Health Statistics (NCHS) conducted a National 2009 H1N1 Flu Survey which was sponsored by the National Center for Immunization and Respiratory Diseases. The one-time survey was a list-assisted random-digit-dialing telephone survey of households. The survey was designed to monitor influenza immunization coverage in the 2009 to 2010 season.

Survey respondents

The target population was persons 6 months or older living in the United States. The data includes surveys from more than 26,000 people.

Overview

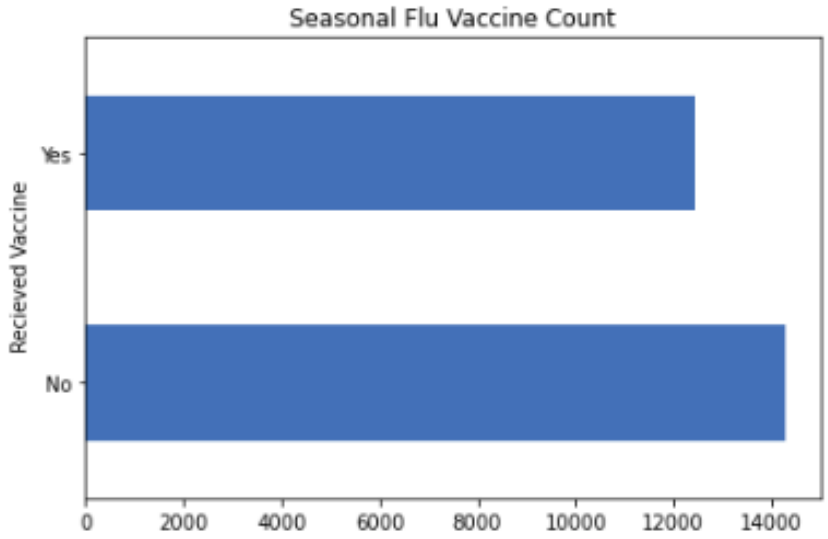
I aim to develop and compare three distinct machine learning models, namely Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, to predict individuals' likelihood of receiving a vaccine. The project will involve preprocessing the dataset, training and tuning the models, and evaluating their performance using appropriate metrics to identify the most effective approach for vaccine distribution prediction.

Dataset Choice

The decision to use this particular dataset was based on both its relevance to public health and its size. It is crucial to investigate the features of individuals who choose to receive the seasonal flu vaccine and those who do not. By analyzing this data, we can predict the likelihood of vaccine compliance and apply this information to other vaccines.

Modeling

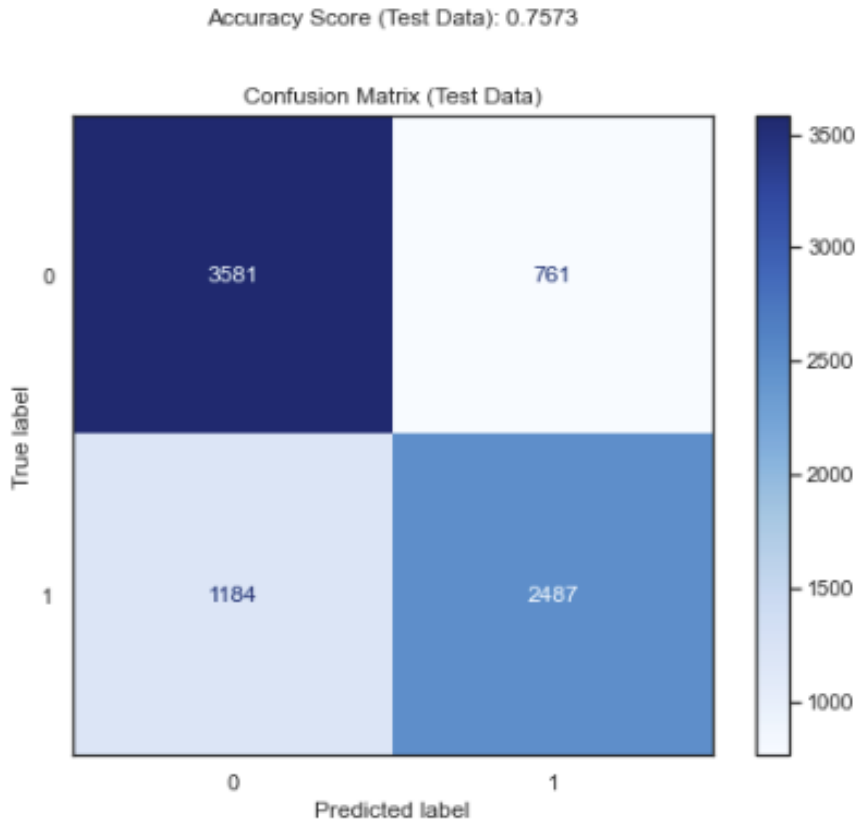
Seasonal Flu Vaccine Count



Two groups were identified: 12,435 individuals who received the flu vaccine and 14,272 individuals who did not receive the flu vaccine. In order to predict the likelihood of individuals receiving the flu vaccine, I utilized accuracy as a metric, as it can be used to measure how often the model correctly predicts whether an individual did or did not receive the vaccine.

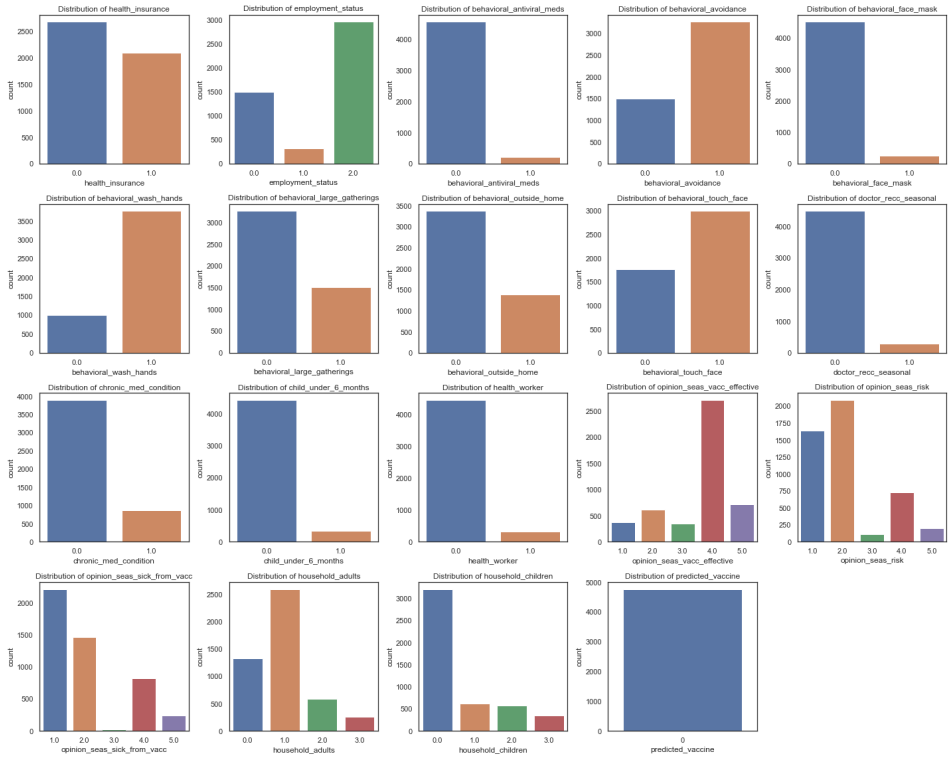
Decision Tree Classifier

```
Classification Report (Test Data):
              precision    recall  f1-score   support

           0       0.75      0.82      0.79      4342
           1       0.77      0.68      0.72      3671

    accuracy                           0.76      8013
   macro avg       0.76      0.75      0.75      8013
weighted avg       0.76      0.76      0.76      8013


Accuracy Score (Test Data):
0.7572694371646075

Accuracy Score (Train Data):
0.7601369423344388
```

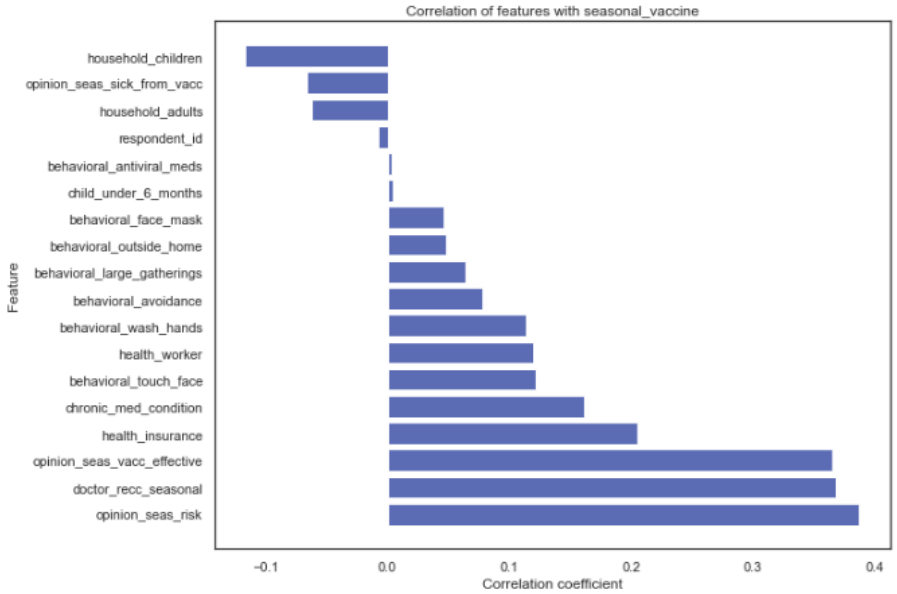


The dataset contained both categorical and numerical features, and the Decision Tree classifier was selected as the final model due to its higher accuracy score compared to the Logistic Regression and Random Forest models. The accuracy score on the test data reflects the model's ability to accurately predict both 0 and 1 classes, and the Decision Tree classifier achieved an accuracy of 75.7% on the test data. The accuracy score on the train data was 76%, and the similarity in accuracy scores between the train and test data indicates that the model is able to generalize well to unseen data.

Evaluation



This visualization displays the distribution of the survey results, revealing that the majority of respondents did not have health insurance. Additionally, it is interesting to note that even health workers at the time did not receive the flu vaccine.



Based on the values in this matrix, we can see that some of the features have moderate to strong positive correlations with each other, such as "opinion_seas_risk", "doctor_recc_seasonal", and "opinion_seas_vacc_effective". This suggests that individuals who perceive a higher risk of the flu, receive a recommendation from their doctor to get vaccinated, and believe that the vaccine is effective are more likely to get vaccinated.

On the other hand, some features have weak negative correlations with each other, such as "household_children" and "opinion_seas_sick_from_vacc". This suggests that individuals who have more children in their household are less likely to believe they will get sick from the vaccine.

Overall, the correlation matrix helps to identify which features may be most important in predicting whether or not someone gets the flu vaccine.

Recommendations

Based on the feature importance results, the top three most important features for predicting whether someone gets the seasonal flu vaccine are:

1. opinion_seas_vacc_effective
2. doctor_recc_seasonal
3. opinion_seas_risk

Therefore, one recommendation would be to focus on improving people's perception of the effectiveness of the vaccine and increasing recommendations from doctors. This could involve public health campaigns and education initiatives to better inform people about the benefits of getting vaccinated and addressing common misconceptions or concerns.

Additionally, the model suggests that age and health worker status are also important factors to consider. Therefore, targeted outreach to older adults and healthcare professionals may also be effective in increasing vaccination rates.

Finally, it's worth noting that some of the other features had very low importance in the model, such as employment status and behavioral habits. While these factors may still be important for individual decision-making, they may not have as much impact on whether someone actually gets vaccinated. Therefore, resources and efforts may be better spent on targeting the factors with higher importance.

Limitations and Next Steps

Data collection was conducted through telephone surveys and could include, limited access to certain populations, non-reponse bias, inaccurate responses and exclusion of non-English speakers. Collecting data through only telephone surveys can limit the sample size and other methods of data collection may need to be considered to minimize these limitations. Despite being collected via telephone surveys, the respondents provided valuable information. To increase flu vaccination rates, the CDC could consider collecting data through additional methods such as online surveys, in-person door-to-door surveys, and by ensuring that surveys are available in multiple languages.

For more information

See the full analysis in the Jupyter Notebook

For additional info, contact Brittney Nitta-Lee at bnittalee@gmail.com

Repository Structure

```
├── .ipynb_checkpoints/
├── Data
├── Images
├── PDFS
├── .DS_Store
├── .gitattributes
├── Notebook.ipynb
└── README.md
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%