dragunat2016 /
CDCH1N1

Code Issues Pull requests Actions Projects Wiki Security Insights

0 stars 0 forks 1 watching Activity

Public repository

main

Branches Tags

dragunat2016 Completed project ... now 8

View code

README.md

CDC Vaccination Status for Seasonal Flu

Overview

The goal of this project is too predict whether people got H1N1 and seasonal flu vaccines using data collected in the National 2009 H1N1 Flu Survey. This is a binary classification problem where we will be investigating if a respondent received the Seasonal flu vaccine.

Repository Structure

- index.ipynb # Jupyter notebook for project
- images # Folder with Images
- training_set_features.csv # CSV with training data
- test_set_features.csv # CSV with testing data
- training_set_labels # CSV with dictionary of columns
- presentation.pdf # High - Level Presentation
- README.md

Business Case

In this study, we will predict whether a participant will get the seasonal flu vaccine. We will optimize on identifying the most amount of people who need a vaccine, at the expense of erroneously identifying people who have already got the vaccine. In data science terms, we will optimize on reducing the amount of false negatives.

The rationale here is that most people who die from the flu are unvaccinated. Ensuring that people are vaccinated will save lives.

Approach

1. Understand the data. Evaluate the columns and features of the data set.
2. Determine if there are data integrity or completeness issues.
3. Fix the data issues in the dataframe.
4. Determine which features to use in the model.
5. Create a baseline model
6. Add to that model.
7. Evaluate the effectiveness of the models and pick a final model to evaluate.

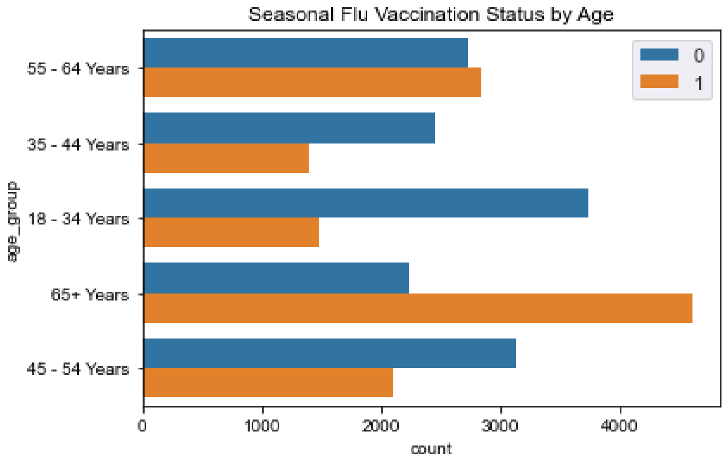
Data Description

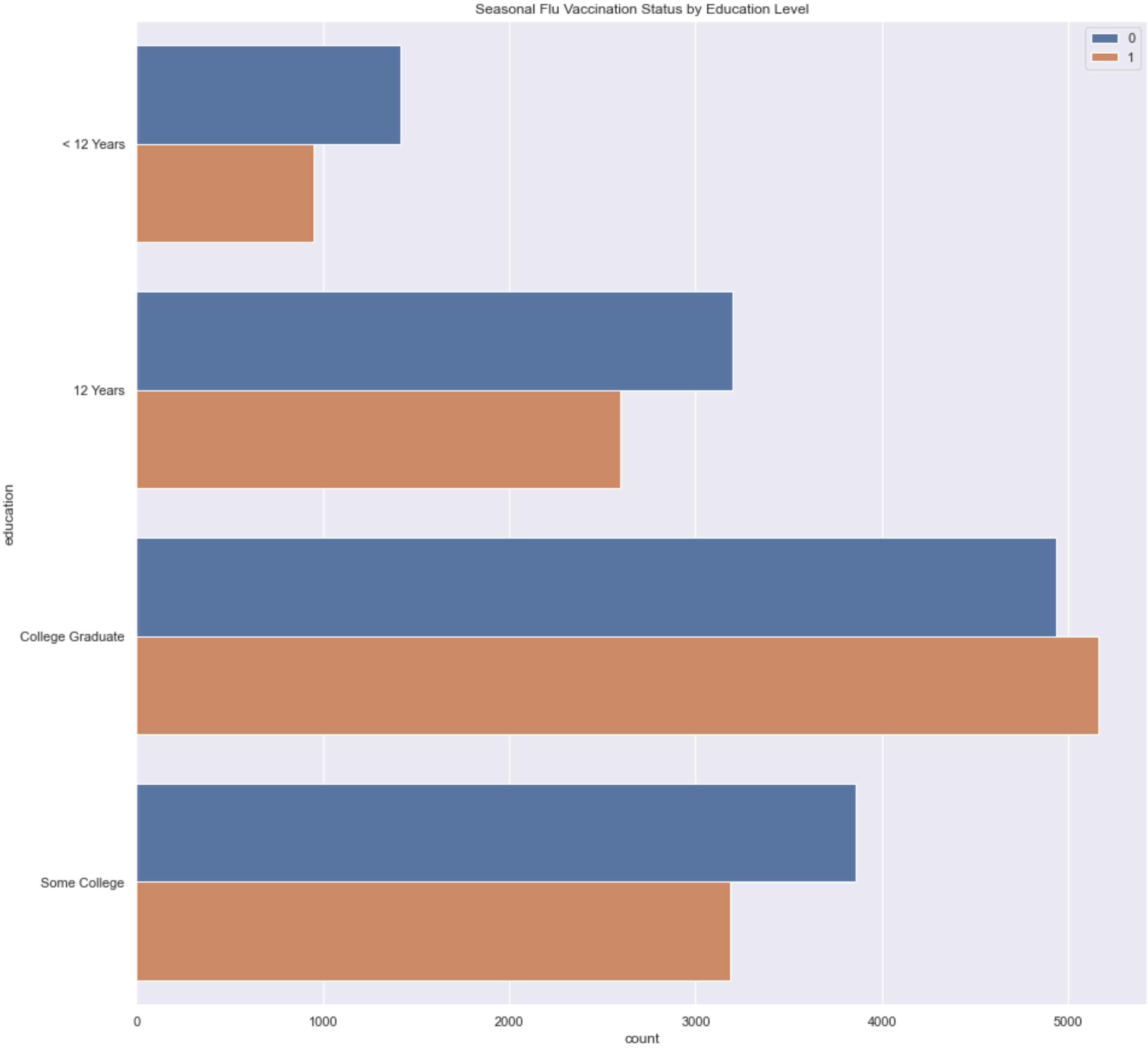
- There are three sets of CSVs provided as part of this study.
- Two of the CSVs are for modeling training, and one is for model testing.
- A train - test split has already been done for us, but we will perform a train - test split on the training data anyway.
- The CSV 'training set features' provides columns we can use to target whether a participant got the Flu Vaccine or not
- The CSV training set labels contains the target column.

Exploratory Data Analysis

Created several slices of the data to explore multiple relationships

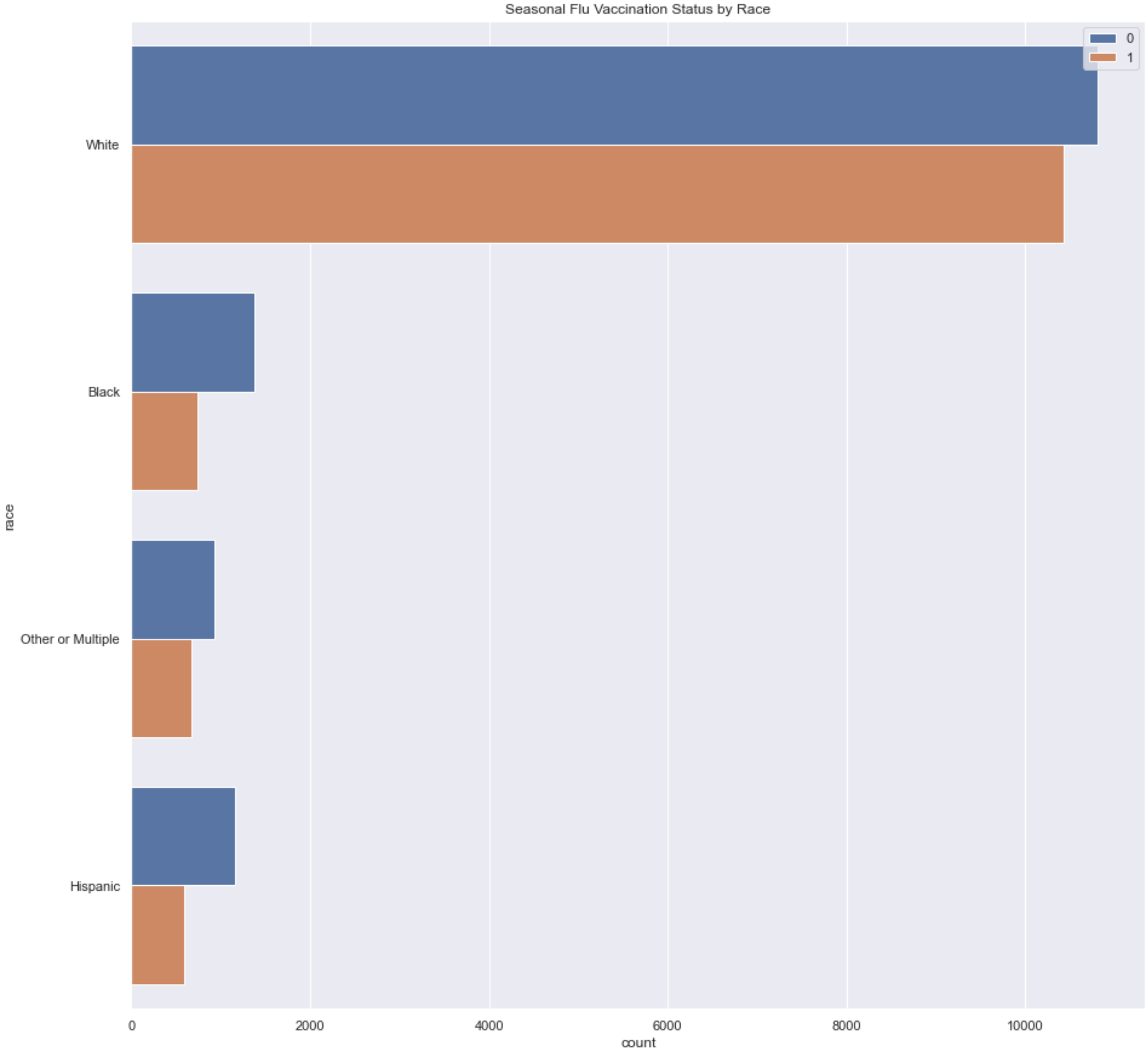
First created a graph comparing the vaccination status of patients based on age



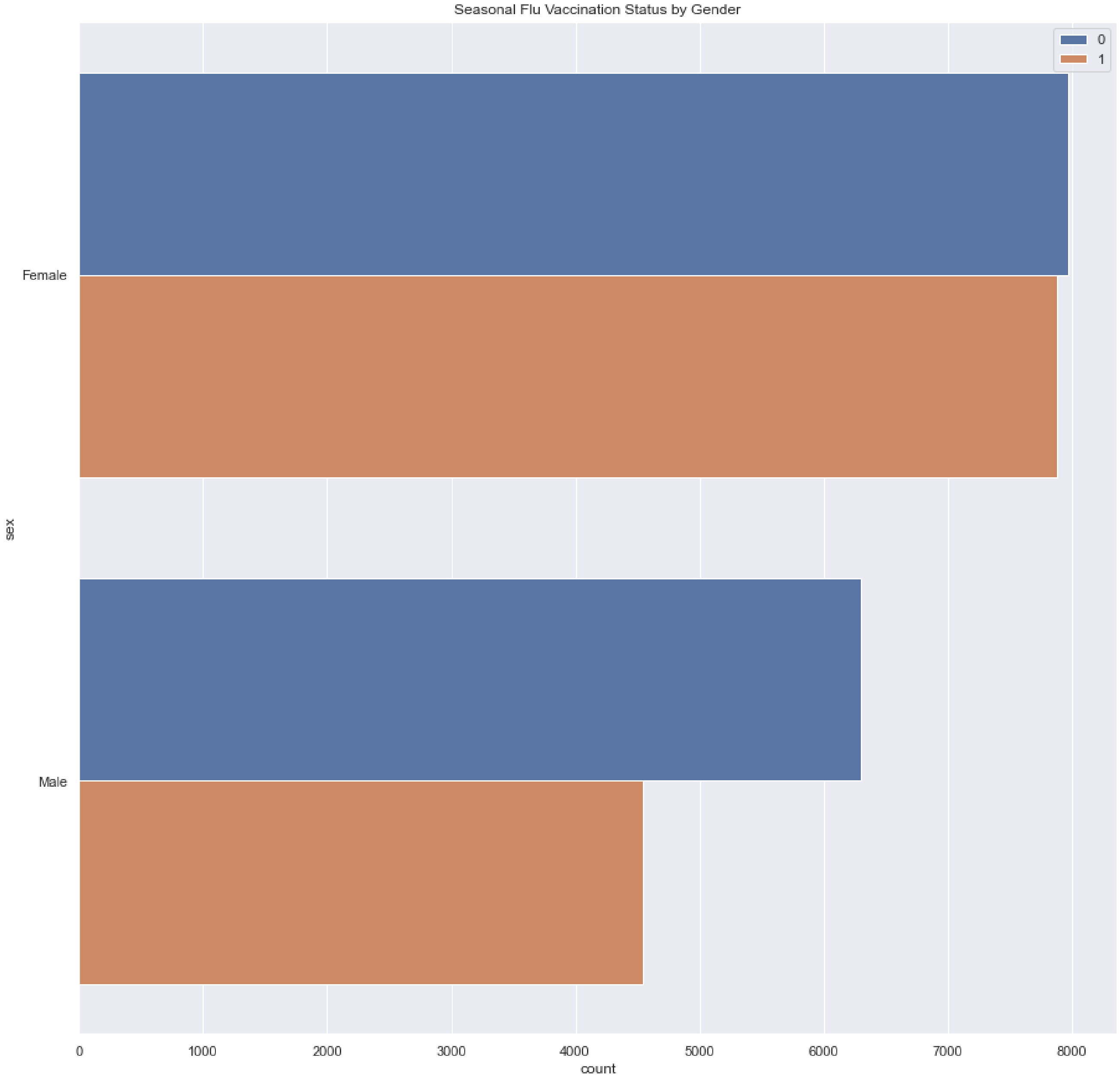


Second compared education level against vaccination status

Compared race to vaccination status



Compared gender to vaccination status



Data Preparation

Here is the process we followed for data preparation

Dropped columns not related to Seasonal: 'h1n1_concern', 'h1n1_knowledge','doctor_recc_h1n1','opinion_h1n1_vacc_effective', 'opinion_h1n1_risk','opinion_h1n1_sick_from_vacc', Drop Columns: health insurance, employment_industry, employment occupation, income_povery drop rows: h1n1_, behavioral_, opinion_h1n1_vacc_effective, opinion_h1n1_risk, opinion_h1n1_sick_from_vacc, household_adults, household_children fill with median: doctor_recc_*, health_worker fill with mode: chronic_med_cond, child_under_6_months, health_worker, education, marital_status, rent_or_own,employment_status

Modeling

Baseline Model

A logistical Regression model will be used as the baseline model. Two forms of the model were evaluated. One that used OHE and one that did not. Data scaled using a standard scaler after these methods. Train-test split occurs after transformation methods to prevent data lost.

Evaluation of Baseline Model

Our baseline model has identified 1648 as true positive results. The model identified the result as positive and its true value was positive.

It had 529 False positive results. Where the model identified them as positive, but these values were actually negative.

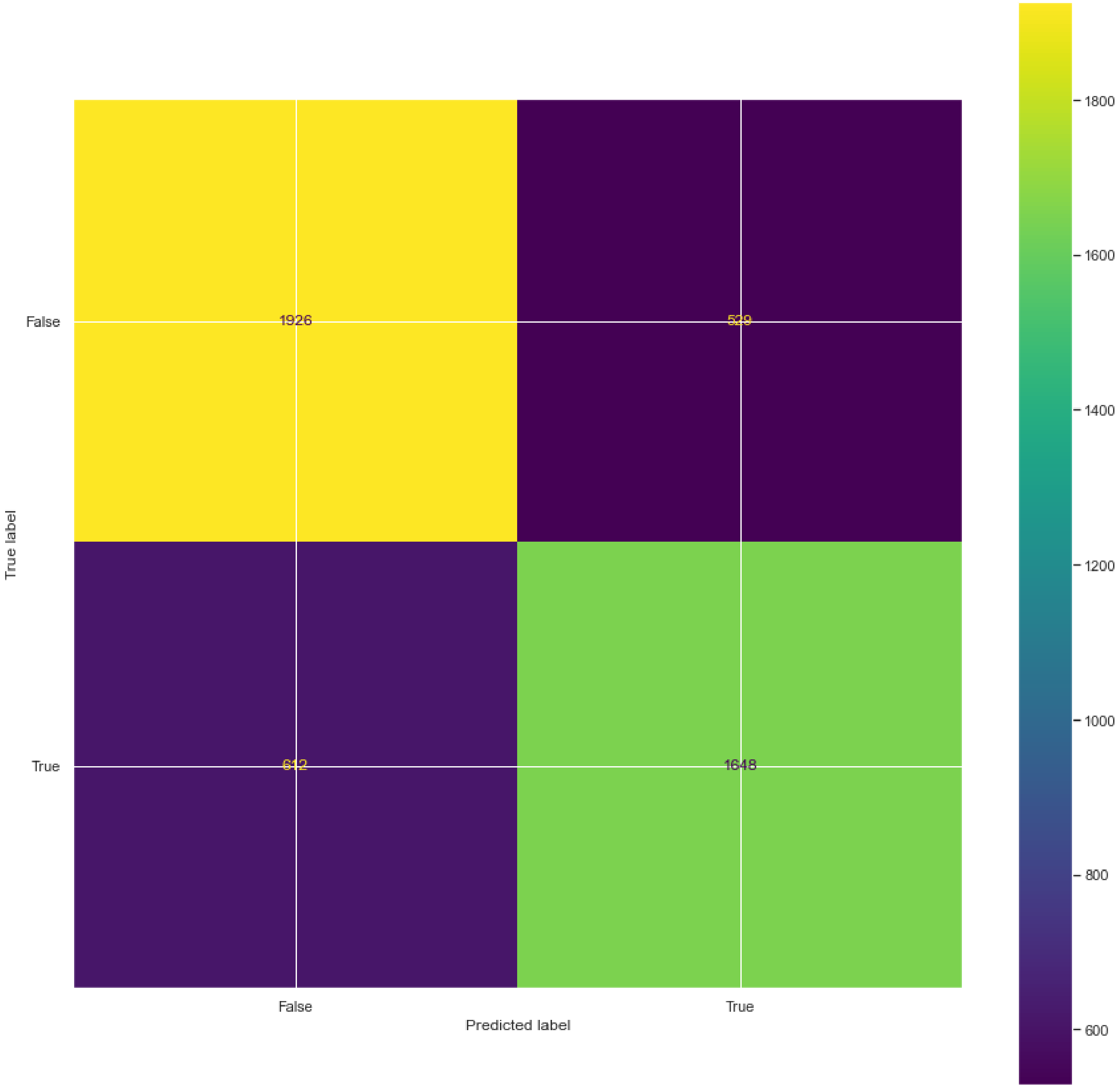
It had 612 False Negative results. Where the model identified them as negative, but they were actually positive.

Finally, it identified 1926 true negatives. These values were negative and the model identified them accurately as negative.

Since we are trying to predict the status of someone needing a vaccine, it's preferable to reduce the number of False negatives. The use of this model would be to determine who needs a vaccination.

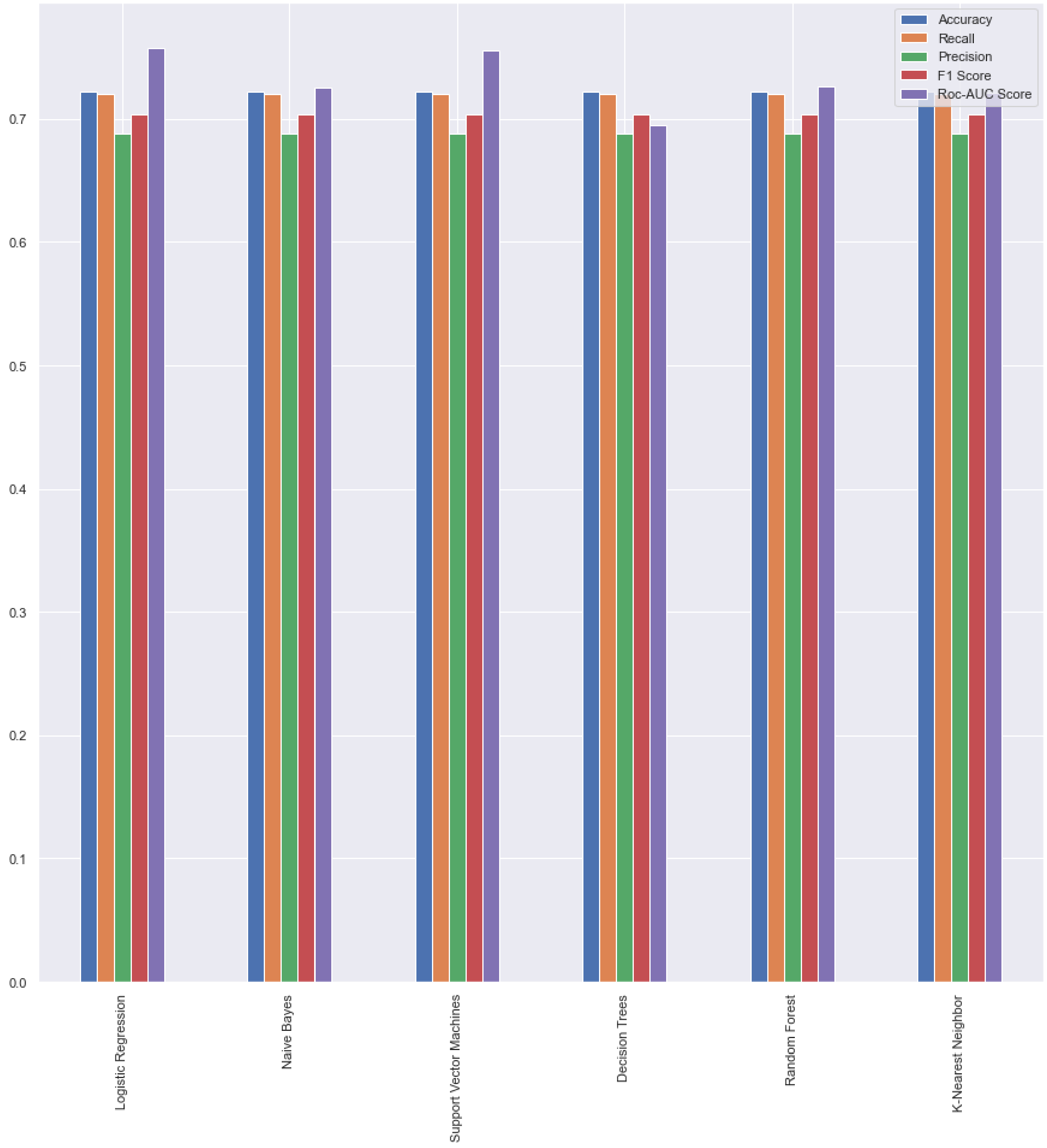
If the model flags a false positive, then a person who got vaccinated, would be pinged again for vaccination.

A false negative would go under the radar, potentially not receive a vaccination, which could be detrimental.



Evaluation of Other Models

Let's evaluate other models to reduce the false negative count of the previous model The following models were evaluated in addition to Logistic Regression: Naive Bayes, Support Vector Machines, Decision Trees, Random Forest, and K-Nearest Neighbor. We calculated the following metrics to evaluate these models against one-another: accuracy, precision, recall, and Roc-AUC score.



Final Model Evaluation

All models have fairly close accuracy, recall, precision, and roc-auc scores. Based on the business problem, we would want to reduce the number of false negatives. A false negative implies that someone was not vaccinated, was predicted to not need a vaccination. That means that this person would be missed if this model would target whom to outreach.

As a result, the logistic regression model and the Support Vector machine model are the top models. Both have accuracy scores of around 76% and recall of 76%.

It seems like the baseline model using logistical regression performed better on the test data with a higher recall and accuracy score. As a result, the final model for this project will be the Logistical regression model.

Future Work

- Consider splitting the data by other aspects like age, race, etc. We optimized on reducing false negatives to save lives. However, the flu is generally fatal to those in certain groups or who have other demographics. We may want to use models that optimize on reducing false negatives in those to save lives. For those who are healthy, getting the flu is not as detrimental. Having too many false positives in those demographics would be a waste of resources. Optimizing on reducing the false positives in those age groups might be more beneficial.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%