H1N1 Vaccination Rate Analysis

By Roseline Maina

Table Of Contents

Business Problem

- Our Client is a medical organization situated in Reno,Nevada. The organization works hand in hand with the government when it comes to the distribution of medical supplies including vaccines within the Reno city in Nevada.
- Using the data on H1N1 vaccine they want you to determine:
    - if more people got vaccinated or not vaccinated
    - which age group is most likely to get vaccinated
    - the impact a doctor's recommendation has on a respondent's decision on whether to get vaccinated or not
    - the distribution of vaccination by gender.
    - They also want you to come up with an algorithm that can predict whether someone got the H1N1 vaccine.
- From this they hope to get a better understanding of how these characteristics have been associated with personal vaccination patterns and will provide guidance for future public health efforts.

MEDICAL
CENTER

Data Understanding

- The data provided had information on over 10,000 respondents where each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey.The data was split into:
  - Training Features : This contained the training input/independent variables. It has 36 columns.
  - Training Labels : This contained the training target/dependant variables. It has 2 columns h1n1_vaccine and seasonal_vaccine of the binary variable types.In our analysis i will be using the H1N1 vaccine target variable only.
  - Test Features: This contained the test input/independent variable. It has 36 columns. But there weren't any target/dependent variable for the test features.

Data Preparation and Exploration

- After loading and familiarizing myself with the data, i prepared the data for analysis by performing the data cleaning.
- I did this by checking the datasets for missing values, duplicates and outliers.
- Next, i used the cleaned datasets to perform the data exploration
- Here, i explored the data by using different visualizations to uncover insights.
- Below, i will give a brief overview on my findings:

# The Distribution of The Target Labels:

Distribution of Target Labels
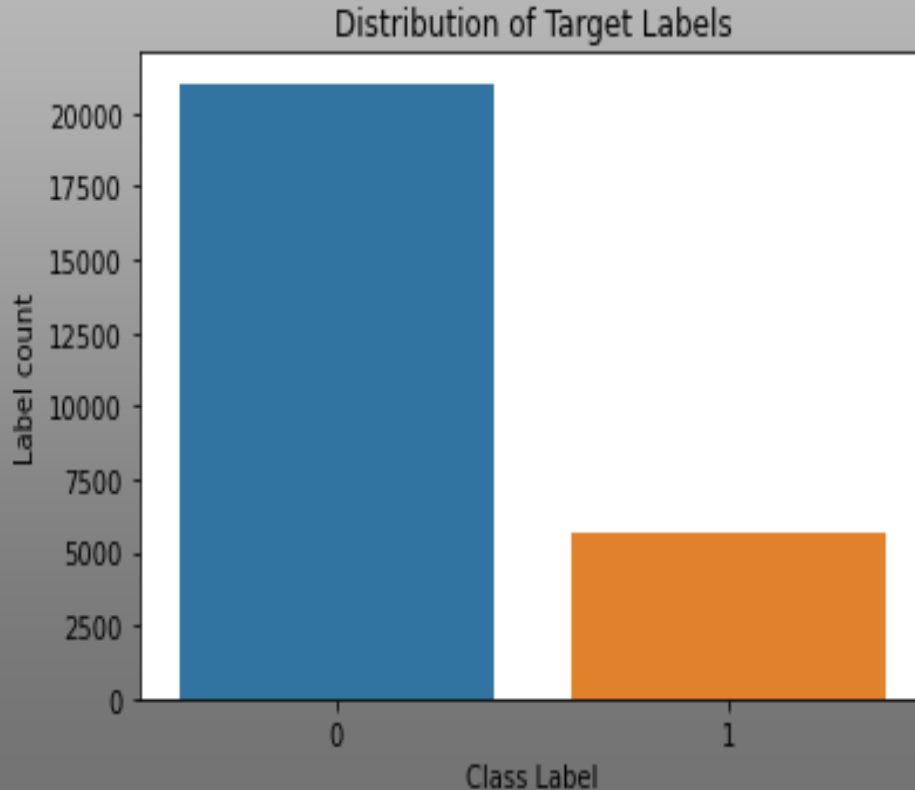
- From the plot on the distribution of the target labels:
  - Class '0' represents unvaccinated respondents and class '1' represents the vaccinate respondents.
  - From this it's clear that class '0' which represents the respondents that did not get the H1N1 vaccine have more data points meaning that most of the respondents were not vaccinated.

The Impact of a Doctor's Recommendation:

Doctor Recommendation:Vaccinated vs Unvaccinated
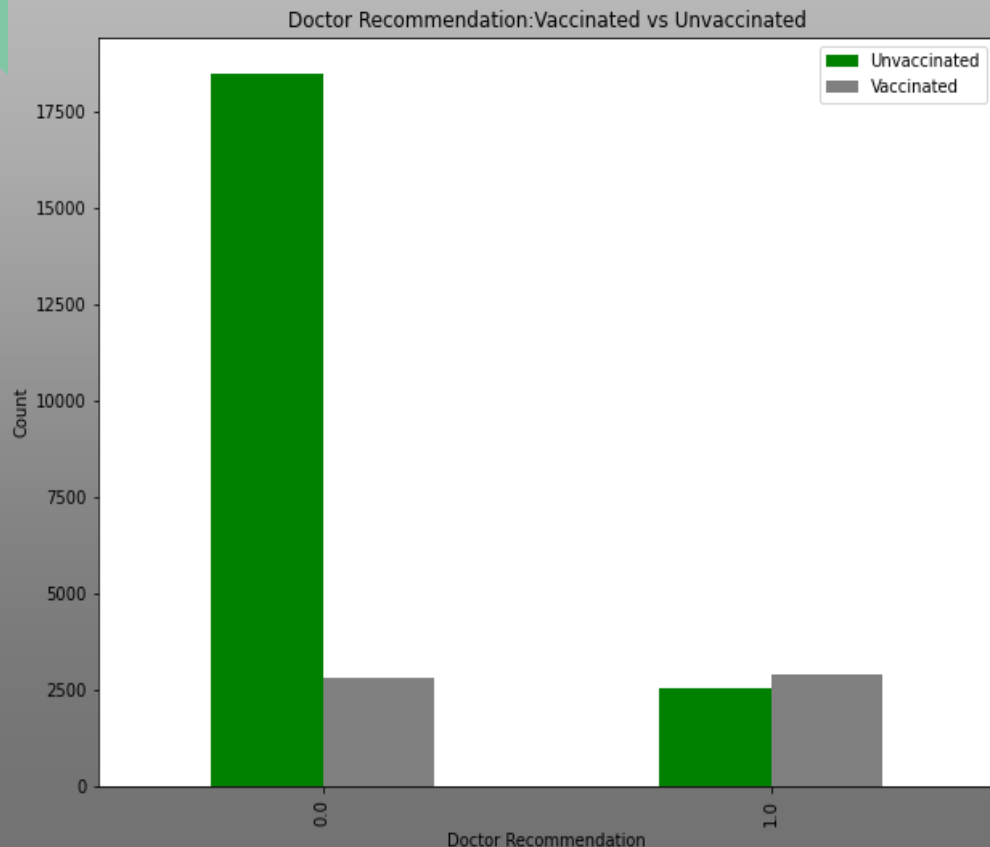
- In the plot, along the x-axis, 0 represents the respondents that did not get a doctor's recommendation and 1 represents the respondents that did.
- From the plot, we see that a doctor's recommendation has little impact on a respondents decision on whether to get vaccinated or not
- This is evident as there is only a slight difference between the vaccinated and unvaccinated respondent's who got a doctor's recommendation

The Age Group most likely to get vaccinated:
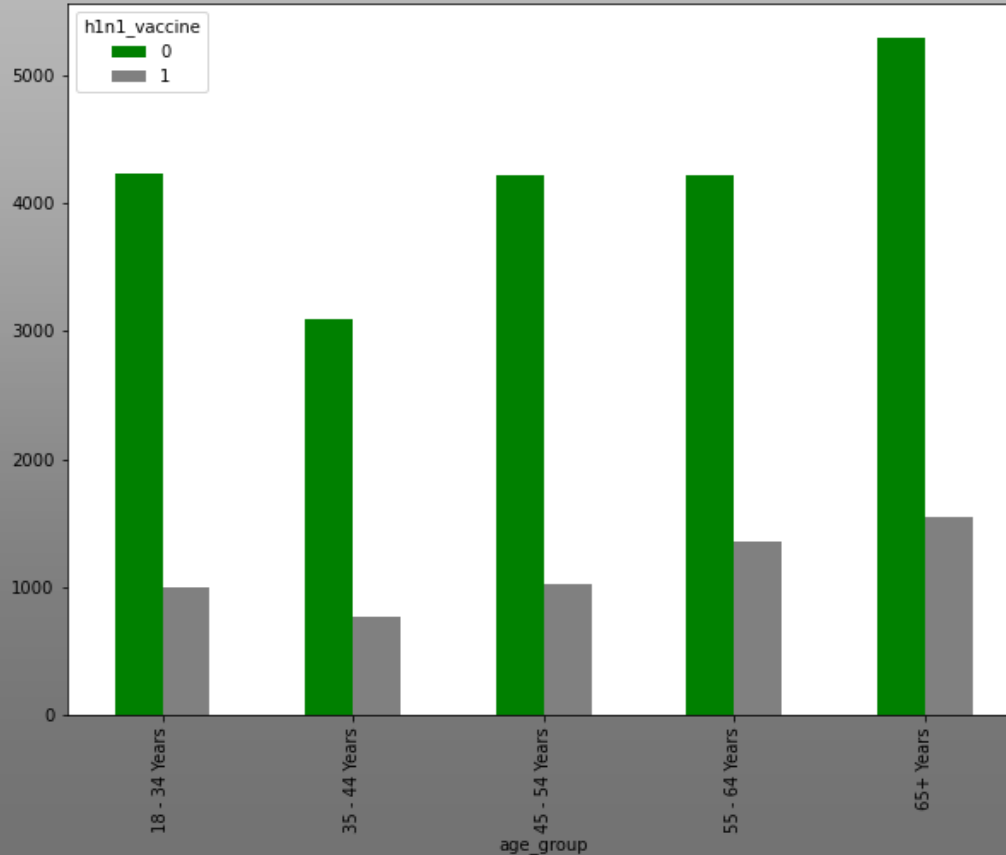
- From the plot:
  - 0 represents the respondents that did not get vaccinated
  - 1 represents the vaccinated respondents

- From this we see that individuals aged 65 years and above are the most likely to go and get vaccinated.

Distribution of Vaccinated by Gender:

Gender:Vaccinated vs Unvaccinated
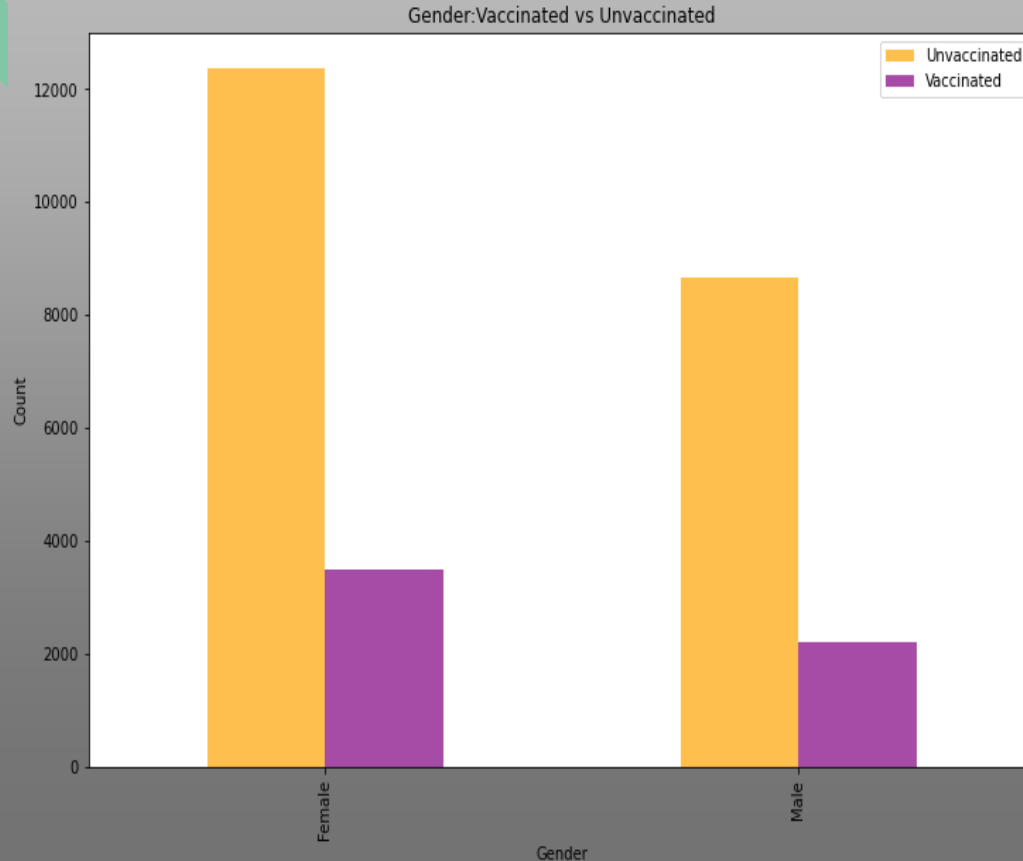
- We see that female respondents have the highest number of unvaccinated respondents when compared to male and at the same time female respondents had the highest number of vaccinated respondents than the male respondents.

Data Modeling and Evaluation

- After performing the exploratory analysis and preprocessing the data , i built and evaluated different classification models in an effort to find the model that will provide the best algorithm for predicting whether someone got vaccinated or not.
- The metrics i used in evaluating the models are
  - the accuracy score which tells us how accurately the model will perform in predicting whether an individual got vaccinated or not
  - The AUC score which tells us how good our model is at distinguishing between the individuals that got vaccinated and those that did not get vaccinated
- Below, i will provide an overview on the final model:
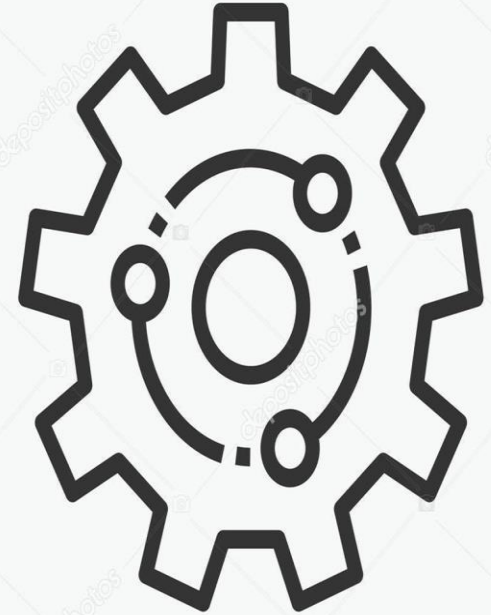
DATA MODELING

Evaluating the Final Model:
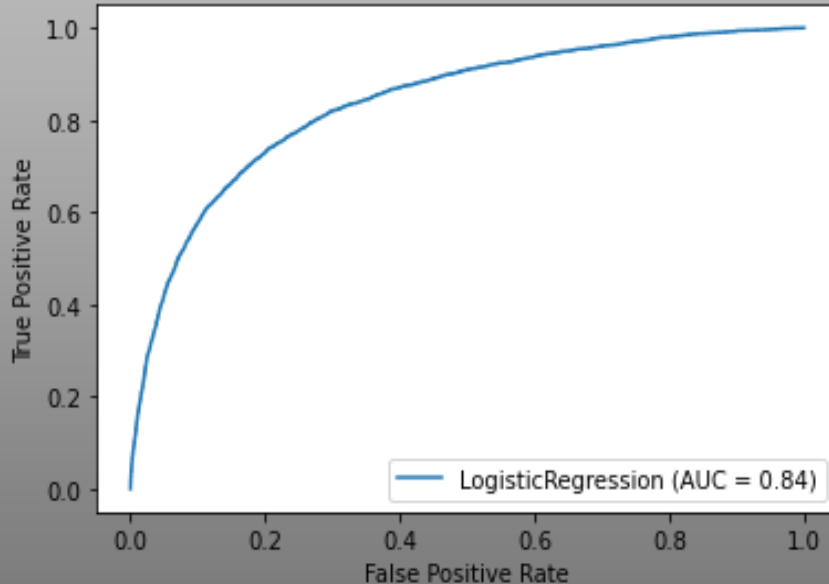
- The dataset used when building the final model was the entire train dataset for both features and labels.
- I used the AUC score when determining which model to fit the final model on.
- It was fitted on the Logistic regression model which provided the best AUC score.
- The plot shows the ROC curve that gives us an AUC score of 84%, this tells us that there is a 84% chance that our model will be able to distinguish between the individuals who got vaccinated and those who did not get vaccinate.
- We also got an accuracy score of 83% which means that our model will get 83% of the predictions (on whether an individual got vaccinated or not vaccinated) right.

Recommendations and Next Steps

- Based on the results from the analysis i'd recommend that:
  - Further investigation is done to understand the underlying reasons that would make an individuals opt not to get vaccinated even though getting vaccinated is beneficial to one's health
  - Further investigation is done to understand why a doctor's recommendation on vaccination does not bear a significant impact on a respondent's choice on whether to get vaccinated or not
- For the next steps:
  - Further exploration should be done to discover whether the general population has any misconceptions or misinformation on vaccinations.

RECOMMENDATIONS

Thank You.