CP-AIML-I24 Capstone – Project 1

Team Members: Asha Nair, Divya, Kunal Kalia, Sri Chitluri

Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines

Problem Statement (Abstract)

Can you predict whether people got H1N1 and seasonal flu vaccines using information they shared about their backgrounds, opinions, and health behaviors?

Our goal is to predict how likely individuals are to receive their H1N1 and seasonal flu vaccines. Specifically, we'll be predicting two probabilities: one for h1n1_vaccine and one for seasonal_vaccine.

Each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey.

—-------------------------------------

Project description and files are from Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines

Expected Outcomes

We expected that individuals are more likely to receive their H1N1 and seasonal flu vaccination when they are more knowledgeable about the vaccines, have behavioral practices to protect health and prevent illness, has a doctor who spoken about the vaccines with them, has previous medical conditions, has any industry knowledge through job, has insurance, has young children in the house, and has employment to pay any costs. We also expect older adults to be more likely to be vaccinated.

Submission Format

From DrivenData problem statement:
The format for the submission
.csv file is three columns: `respondent_id`, `h1n1_vaccine`, and `seasonal_vaccine`.

The predictions for the two target variables should be *float* (`float64`) **probabilities** that range between `0.0` and `1.0`. We need to build a model to predict two numeric probabilities for these target variables:

- `h1n1_vaccine`
- `seasonal_vaccine`

Because the competition uses `ROC AUC` as its evaluation metric, the values you submit must be the **probabilities that a person received each vaccine**, not binary labels.

As this is a multi-label problem, the probabilities for each row do not need to sum to one.

Binary labels are provided in the `training_set_labels` dataset. The labels are the values we need to predict. The `training_set_features` dataset has the features, the independent variables used to predict the probability.

Solution Approach

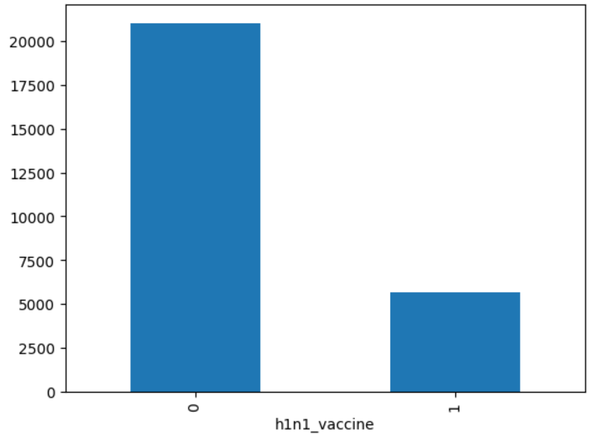
Data Extraction - Fetch the data from the source

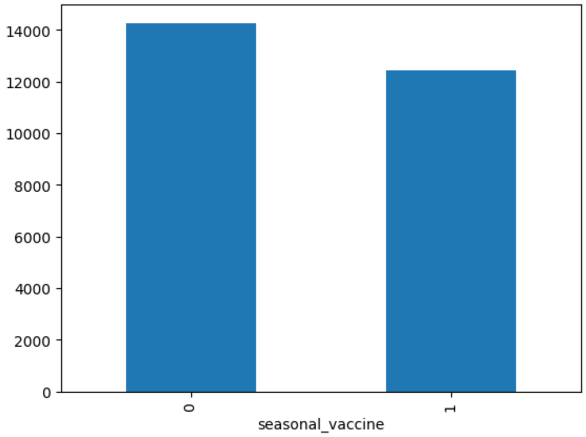
Several libraries were imported, for example: `numpy`, `pandas`, `matplotlib`, `seaborn`, and `sklearn`. The data from the training *.csv* spreadsheets was imported into `Panda` data frames. The data frames made include one for *features*, one for *labels*, and one with both labels and features combined. The index column of the frames is by `respondent_id`, as this is a unique value. The data frames were examined for the number of columns and rows (dimensions of the dataset). This gives us an idea of how many samples (observations) and features (variables) we have. We see that each row corresponds to a submission, with 26,707 submissions for 35 features (which are the columns of the *features* data frame). From the *label* data frame, we see the same number of submissions for the two target variables.

Data Pre-process

The data type for each feature was examined and if there were any null values in the data. Any variables that have the data type of "*object*" will need to be encoded into an "*int*" data type, for easier analysis by the ML algorithm. The dataset was checked for any duplicates, of which there were none.

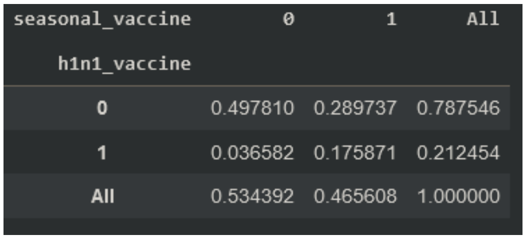
The data for the target variables was explored. From the results, we see that for the seasonal vaccine, around ~46% people received it. For the H1N1 vaccine, only around ~21% people have received it. Classes are imbalanced for the H1N1 target class. (In this dataset: `0 = not received vaccine` & `1 = received the vaccine`)





We also checked if there was any influence between the two target variables; as in if one vaccine was received, was the other vaccine also received?

- Looks like, the people who got the H1N1 vaccine → most also got the seasonal vaccine.
- The people who got the seasonal vaccine → more did not get the H1N1 vaccine.
- However, people who did not get the seasonal vaccine, more than likely also did not get the H1N1 vaccine.



seasonal_vaccine	0	1	All
h1n1_vaccine			
0	0.497810	0.289737	0.787546
1	0.036582	0.175871	0.212454
All	0.534392	0.465608	1.000000

So for the class imbalance for H1N1 vaccine, we need to take this into account when splitting our train set into train/validation.

There were a few features that had the most missing values: `employment_occupation`, `employment_industry`, `health_insurance`, and `income_poverty`. These are features that survey participants did not want to answer (perhaps the question was uncomfortable to answer), and should be kept in mind when looking over the data.

Two lists were made, an array for the category data and an array for numerical data, by checking if the data type was an *object* or not for that variable. Then we encoded the category data variables to be an *int* data type in all the data frames (though we encoded some features later on, just so the data graphs look nicer).

These are the feature variables:

demographic

- age_group
- race

These are the independent variables:

- demographic
- socio-economic
- employment

- sex
- marital_status

socio-economic

- health_insurance
- education
- income_poverty • rent_or_own

employment

- employment_status
- employment_occupation
- employment_industry
- health_worker

health

- h1n1_ variables (2)
- doctor_ variables (2)
- chronic_med_condition

household & family

- child_under_6_months
- household_adults
- household_children

geographical

- census_msa
- hhs_geo_region opinion
- opinion_ variables (6) behavioral
- behavioral_ variables (7)

- geographical
- health
- household & family

These are the target variables:

targets:

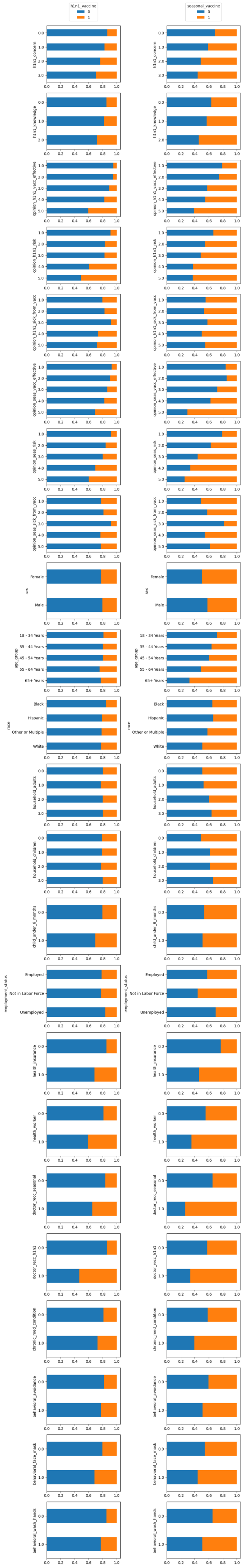
- h1n1_vaccine (yes/no)
- seasonal_flu vaccine (yes/no)

others:

- behavioral
- opinion

Data Analysis and Visualization

Then we visualized a few of the features against the H1N1 vaccine and seasonal vaccine target variables. Doing this comparison graphs by response count was not helping identify trends. Instead we figured out we need to change the data in the bar graphs to show the percentage of how many were vaccinated vs not vaccinated out of the total people for each level of said feature. The bars are stacked because the percentages total out to 1.0.


h1n1_vaccine
0
1
seasonal_vaccine
0
1
h1n1_concern
0.0
1.0
2.0
3.0
h1n1_knowledge
0.0
1.0
2.0
opinion_h1n1_vacc_effective
1.0
2.0
3.0
4.0
5.0
opinion_h1n1_risk
1.0
2.0
3.0
4.0
5.0
opinion_h1n1_sick_from_vacc
1.0
2.0
3.0
4.0
5.0
opinion_seas_vacc_effective
1.0
2.0
3.0
4.0
5.0
opinion_seas_risk
1.0
2.0
3.0
4.0
5.0
opinion_seas_sick_from_vacc
1.0
2.0
3.0
4.0
5.0
sex
Female
Male
age_group
18 - 34 Years
35 - 44 Years
45 - 54 Years
55 - 64 Years
65+ Years
race
Black
Hispanic
Other or Multiple
White
household_adults
0.0
1.0
2.0
3.0
household_children
0.0
1.0
2.0
3.0
child_under_6_months
0.0
1.0
employment_status
Employed
Not in Labor Force
Unemployed
health_insurance
0.0
1.0
health_worker
0.0
1.0
doctor_recc_seasonal
0.0
1.0
doctor_recc_h1n1
0.0
1.0
chronic_med_condition
0.0
1.0
behavioral_avoidance
0.0
1.0
behavioral_face_mask
0.0
1.0
behavioral_wash_hands
0.0
1.0
0.0
0.2
0.4
0.6
0.8
1.0

The features asking about knowledge or opinion of the vaccines show good indications for the target variables.

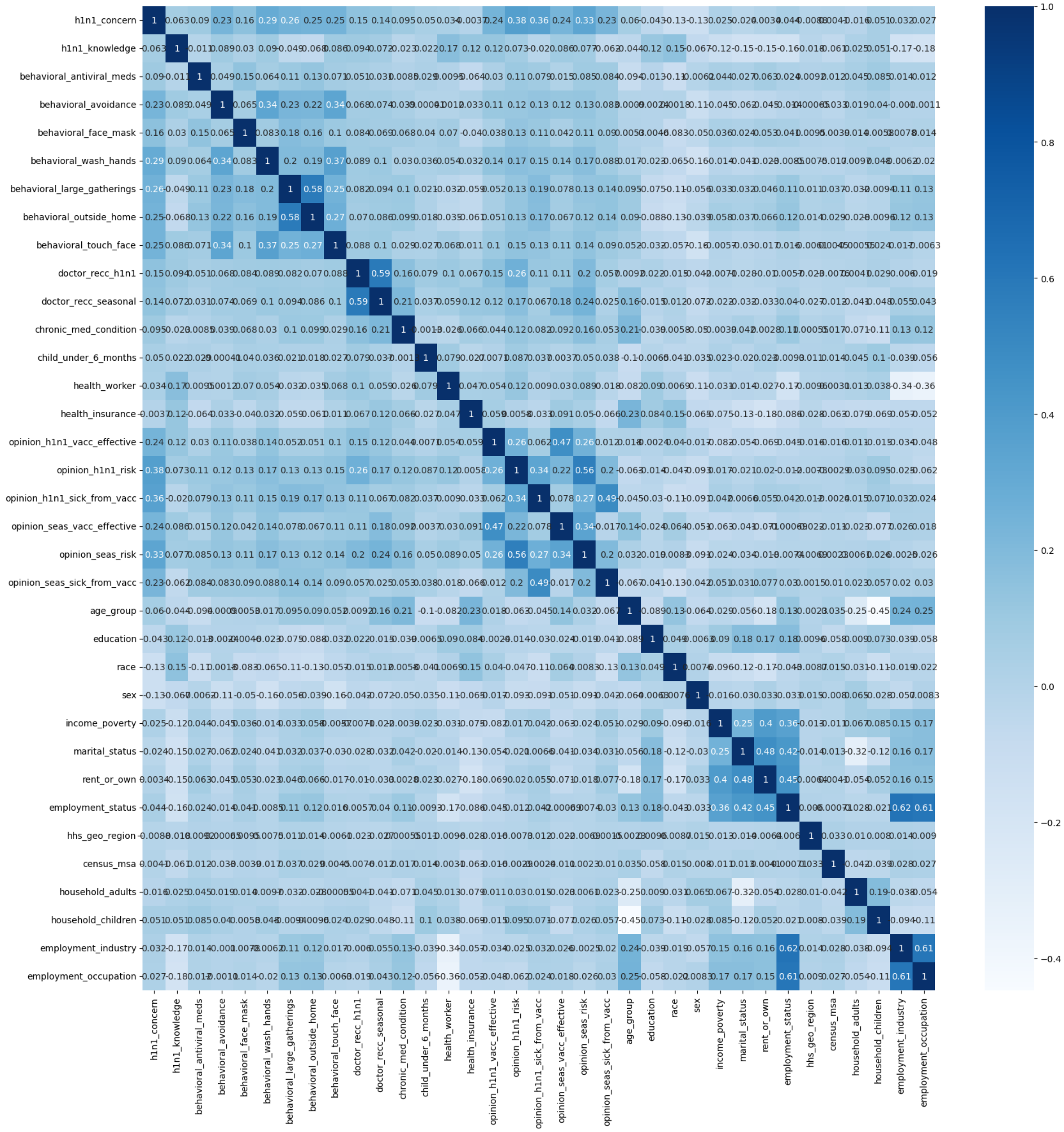
- The more knowledge of H1N1: more vaccinated for H1N1 and seasonal
- H1N1/seasonal effectiveness opinion, think more effective: more percentage vaccinated for H1N1 and seasonal
- The opinion of higher the risk of getting sick from H1N1/seasonal flu without vaccine: more percentage vaccinated for H1N1 and seasonal
- The opinion of more worry of getting sick from the H1N1/seasonal vaccine: not all that clear cut, though the median worry level from getting sick from either vaccine has more people not vaccinated compared to the other levels.
- Demographic features show more trends with the seasonal vaccine: the older the person is, more vaccinated; for race, white and other races, more vaccinated (outside historical context: African Americans are more skeptical of healthcare due to previous medical tragedies that happened to African Americans, thus this is understandable); sex of the person, around the same percentage, with just more female persons having more vaccination.
- Number of adults and children in the house seem to not really show any trend. However, people with regular contact with children under 6 months are more likely to be vaccinated for H1N1 and seasonal flu.
- Those who have health insurance, are more vaccinated on H1N1 and seasonal flu. Those who are healthcare workers, are more vaccinated for both.
- Those who were recommended by their doctor for either H1N1 or seasonal flu vaccine, are more vaccinated for H1N1 and seasonal. As seen before, those recommended the seasonal vaccine by their doctor, not many also vaccinated on H1N1. But those recommended the H1N1 vaccine by their doctor, many also vaccinated for seasonal flu.
- For the behaviors, the more positive behavior, the more likely to be vaccinated for both overall.
- Those with chronic medical conditions, more likely to be vaccinated for both, more so the seasonal flu vaccine than the H1N1 flu vaccine.

Feature scaling and Selection

Dimensionality Reduction: We have 35 features in our dataset, and have considered feature selection to improve model performance. Note: Consider PCA if features are highly correlated or use feature selection techniques to retain the most relevant features.

Before heading with ML algorithms, we have performed feature scaling to the columns since features have different units or ranges. We also did label encoding on the categorical data.

Result from correlation heat map:



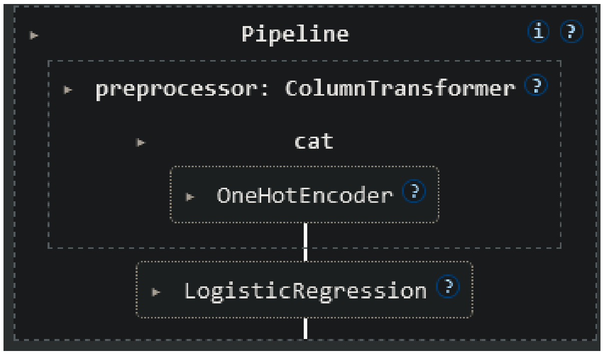
Model: Creation, Selection, Training, Validation, Prediction

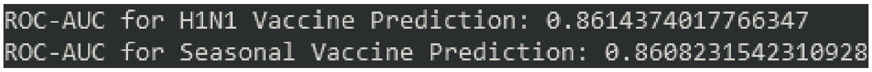
We did model exploration on a separate code file, to test which type of model gives the best performance for the data set. The models tested were: using binary relevance (logistic regression, random forest), using classifier chains (logistic regression, decision tree), and using label powerset (naive bayes). Looking at the accuracy and precision scores for the models tested, we decided to go with Logistic Regression as our model for our final submission file.

This is for the *Logistic Regression* model creation, training, validation, and testing.

We split the label dataset into `h1n1_vaccine` and `seasonal_vaccine` so the model can better learn the relationship between each target variable and the features. We did one-hot encode the feature variables to be more ML algorithm friendly. Then built a pipeline model for the h1n1 vaccine prediction and for seasonal vaccine prediction.

The dataset was split into training and validation for the model training and performance validation. This way, we can check beforehand how the model will do with the test dataset. Fit the training data into the appropriate pipeline model and then made predictions (probabilities) on the validation dataset. The models were evaluated using `ROC-AUC` using the validation set and the probabilities calculated. The `ROC-AUC` metric is used to evaluate model performance on the validation set. `ROC-AUC` can handle imbalanced classes. These are the `ROC-AUC` scores for the 2 models:

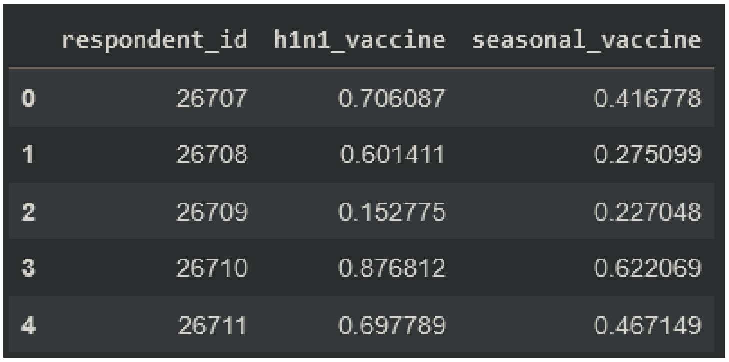




```
ROC-AUC for H1N1 Vaccine Prediction: 0.8614374017766347
ROC-AUC for Seasonal Vaccine Prediction: 0.8608231542310928
```

The models have performed as well as they could, close to .90.

Now the test dataset was loaded into its own data frame, and each vaccine pipeline model was directed to make probabilities using the test set features. The result was loaded into a submission *.csv* file and downloaded. The resulting table shows the probability for the person obtaining each vaccine using the *features* on the testing data set. Here is the `.head()` of the submission results using logistic regression:



	respondent_id	h1n1_vaccine	seasonal_vaccine
0	26707	0.706087	0.416778
1	26708	0.601411	0.275099
2	26709	0.152775	0.227048
3	26710	0.876812	0.622069
4	26711	0.697789	0.467149