H1N1 Acceptance Rates Exploratory Analysis

Freddy Abrahamson

Mar27th, 2022

Outline

- Business Problem
- Data Understanding
- Explaining the Metric
- Model Recommendation
- Feature Recommendation

Business Problem

As a consequence of the Covid 19 pandemic, there is a renewed interest in the vaccination rates for the seasonal flu. We will be performing an exploratory analysis to identify who is more likely to take the seasonal flu vaccine.

To that end, we will go over the best metric to use, the most effective predictive model, and feature Importance.

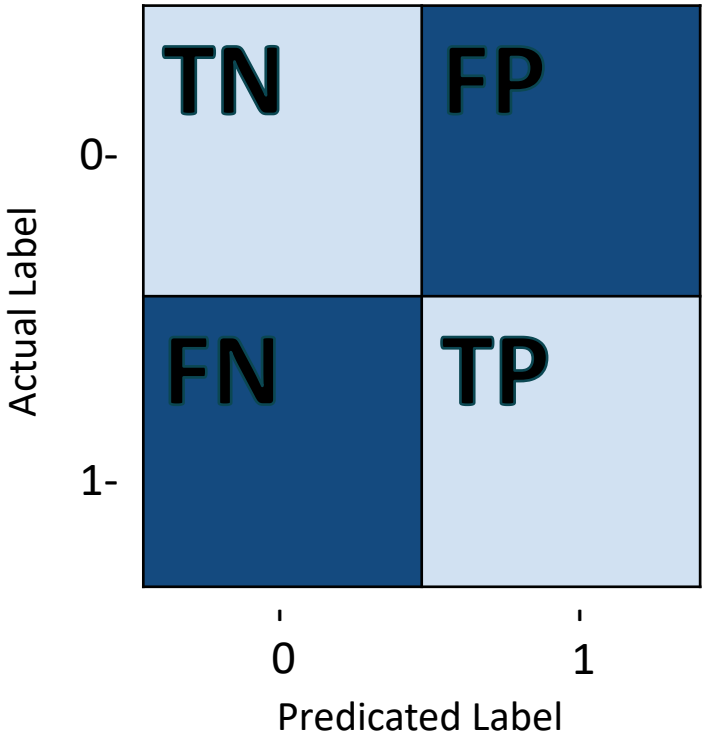
Data Understanding

- The data comes from the National 2009 H1N1 Flu Survey, in the form of a ‘csv’ file.
- The data is comprised of 36 columns, including the target variable, and 26,707 rows.
- Each row represents a respondent in the survey.
- It is an imbalanced dataset with the majority to minority class ratio at approximately 4:1 .
- The survey is primarily comprised of binary and multiple choice questions such as:

h1n1_knowledge : 0 = No knowledge 1 = A little knowledge 2 = A lot of knowledge
health_worker : 0 = no 1 = yes

The Metric: F1 Score Explained

- We will use a predictive model to determine whether the respondent :
  - 1. will take the h1n1 vaccine
  - 2. will not take the h1n1 vaccine
- A model that predicts one of two possible outcomes is called a binary predictive model.
- The predictions made by this model can be put into 1 of 4 categories :
  - 1. true positive: The respondent was predicted to take the vaccine and actually took the vaccine
  - 2. true negative: The respondent was predicted to not take the vaccine and did not take it.
  - 3. false positive: The respondent was predicted to take the vaccine, but did not take it.
  - 4. false negative: The respondent was predicted to not take the vaccine, but did take it.



The Metric: F1 Score Explained II.

- In our use case, which is exploratory analysis, we simply want to predict the information as accurately as possible.

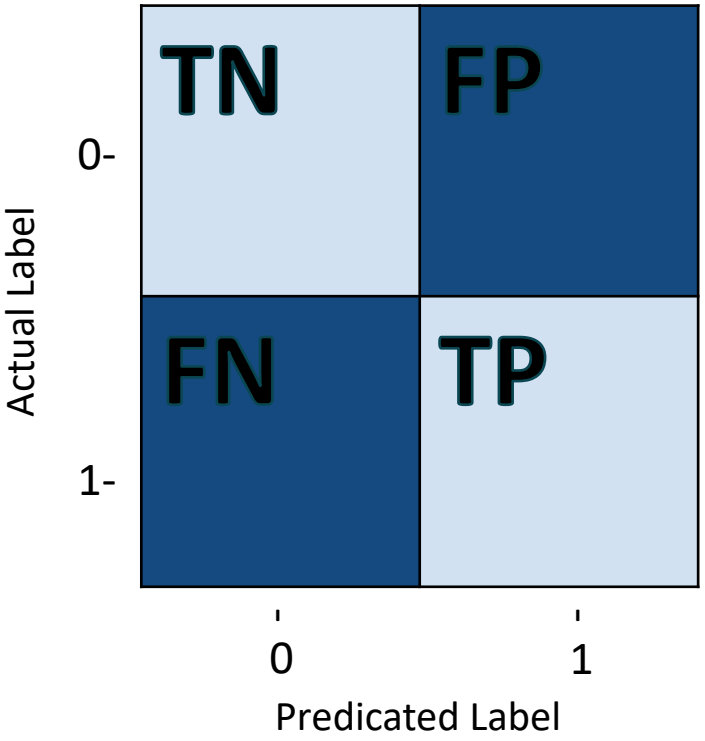
$$accuracy = \frac{number\ of\ correct\ predictions}{number\ of\ total\ predicitions}$$

- Using a metric such as accuracy, for example, that calculates the number of correct predictions made, divided by the number of total predictions, can be misleading when working on an imbalanced data set.
- Let us say for example, that we want to create a model that predicts how many boys playing baseball as high school seniors, will make it into the MLB. For argument's sake, let us put the actual percentage as 1%, or (1/100). If we created a model that did nothing more than simply predict that nobody ever made it into the MLB, the accuracy score of that model would still be 99%.
- This example illustrates how on an imbalanced data set, how even a 'useless' model, that does nothing more than choose the more likely possibility all the time, can still produce a high accuracy score.

The Metric: F1 Score Explained III

- This imbalance is what makes the f1 score metric a very appealing option. It takes data imbalance into consideration by penalizing false positives and false negatives, thereby, at the same time, rewarding true positives, and true negatives.

- The f1 metric not only rewards the model for correct predictions, but is not ‘misled’ by imbalanced data sets.



Model Recommendation

model: Random Forest

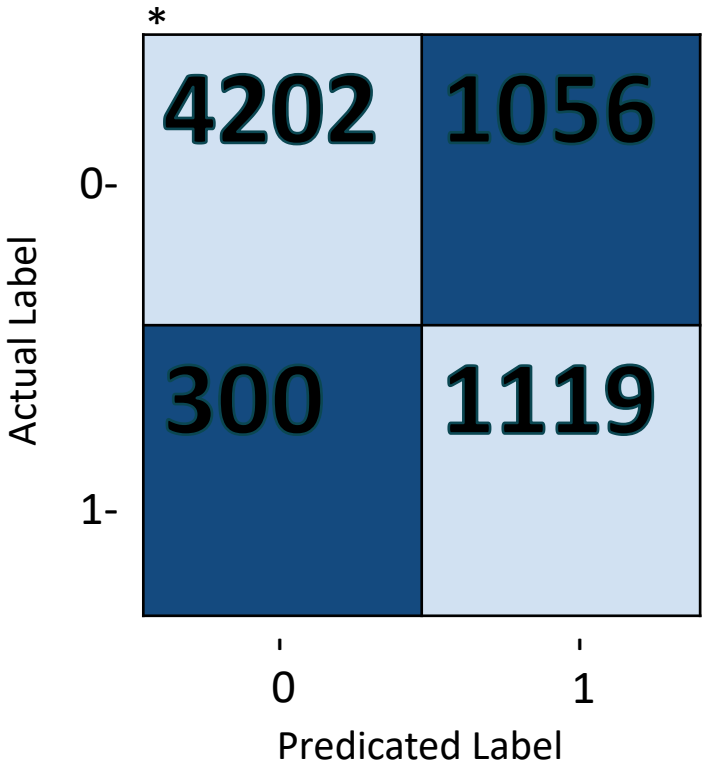
metric: F1 score

F1 score: .65 (with optimized threshold at ~ .57)

model parameters:

class_weight : 'balanced'
criterion : 'gini'
max_depth : 5
max_features : 32
min_samples_split : 2
n_estimators : 150

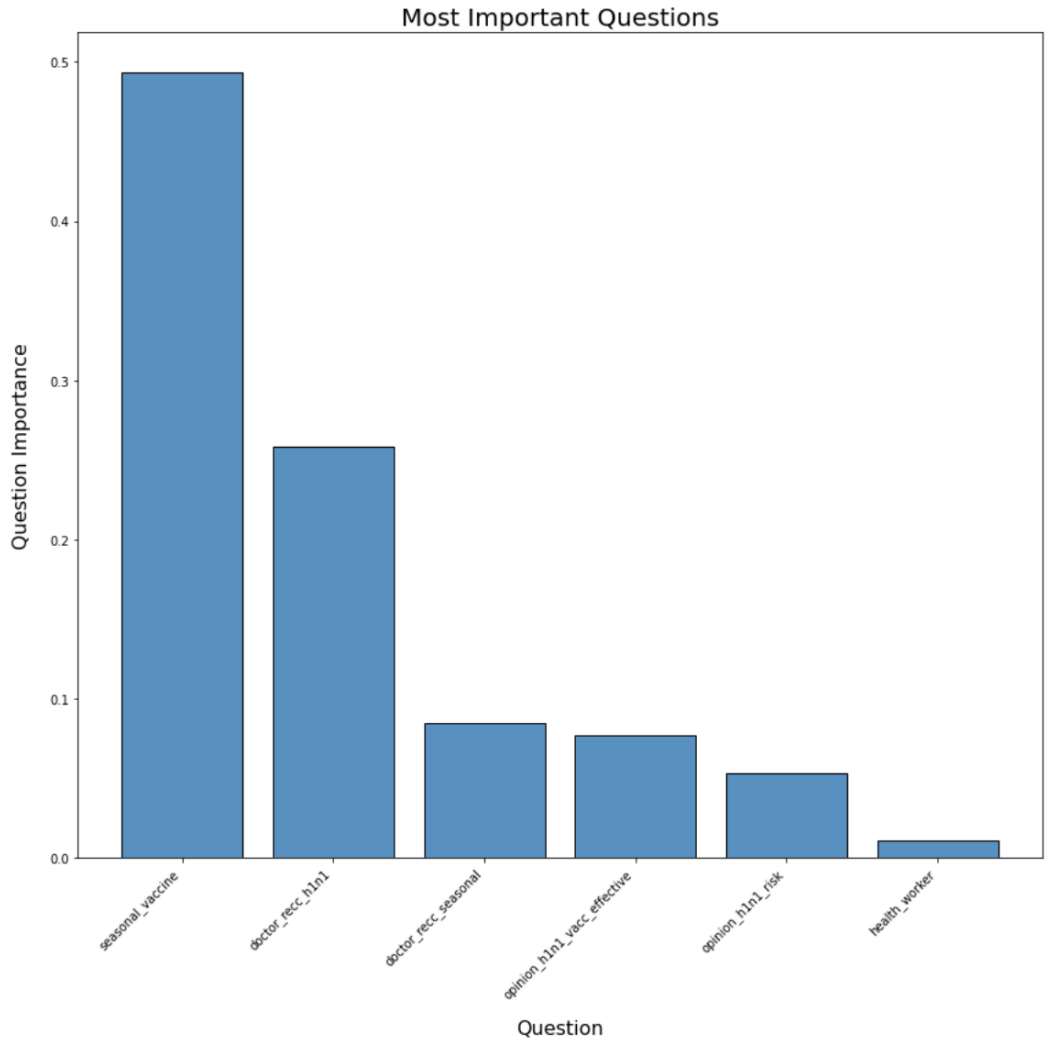
how to handle class imbalance : SMOTE(sampling_strategy=0.40)



*with .5 threshold at .5

Feature Recommendation

- The feature importance of a model is a metric that measures how useful a feature is to the model in making its predictions.
- A feature importance score ranges from 0 to 1, with the sum of all feature importances being equal to one.
- From the chart we can see that the 6 features in this plot account for nearly 98% of the total feature importance.
- I highly recommend that a special emphasis is placed on these features when acquiring new information.



Any Questions?

Thank You