PREDICT PROBABILITY OF RECEIVE H1N1 VACCINE

Final Project Report of

General Business 656 - Machine Learning for Business Analytics

By Sang Lu (Sandra)

Instructor: Dr. Daniel Bauer

December 19, 2021

Background

Beginning in spring 2009, a pandemic caused by the H1N1 influenza virus, colloquially named "swine flu", swept across the world. Researchers estimate that in the first year, it was responsible for between 151,000 to 575,000 deaths globally. A vaccine for the H1N1 flu virus then became publicly available in October 2009.

In late 2009 and early 2010, the United States conducted the National 2009 H1N1 Flu Survey. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, in conjunction with questions about themselves. These additional questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission. A better understanding of how these characteristics are associated with personal vaccination patterns can provide guidance for future public health efforts.

Problem Statement

The competition named *Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines* is conducted by DrivenData. In this competition, our goal is to predict how likely individuals are to receive their H1N1 vaccines using the information they shared about their background, opinions, and health behaviors.

Data Description

The data for this competition comes from the National 2009 H1N1 Flu Survey (NHFS). Each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey.

A list of features in the dataset as below.

- h1n1_concern - Level of concern about the H1N1 flu. (0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.)
- h1n1_knowledge - Level of knowledge about H1N1 flu. (0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.)
- behavioral_antiviral_meds - Has taken antiviral medications. (binary)
- behavioral_avoidance - Has avoided close contact with others with flu-like symptoms. (binary)
- behavioral_face_mask - Has bought a face mask. (binary)
- behavioral_wash_hands - Has frequently washed hands or used hand sanitizer. (binary)
- behavioral_large_gatherings - Has reduced time at large gatherings. (binary)
- behavioral_outside_home - Has reduced contact with people outside of own household. (binary)
- behavioral_touch_face - Has avoided touching eyes, nose, or mouth. (binary)
- doctor_recc_h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
- chronic_med_condition - Has any of the following chronic medical conditions: asthma or another lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary)
- child_under_6_months - Has regular close contact with a child under the age of six months. (binary)
- health_worker - Is a healthcare worker. (binary)
- health_insurance - Has health insurance. (binary)
- opinion_h1n1_vacc_effective - Respondent's opinion about H1N1 vaccine effectiveness. (1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.)
- opinion_h1n1_risk - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine. (1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.)
- opinion_h1n1_sick_from_vacc - Respondent's worry of getting sick from taking H1N1 vaccine. (1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.)
- age_group - Age group of respondent.
- education - Self-reported education level.
- race - Race of respondent.
- sex - Sex of respondent.
- income_poverty - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- marital_status - Marital status of respondent.
- rent_or_own - Housing situation of respondent.
- employment_status - Employment status of respondent.
- household_adults - Number of *other* adults in household, top-coded to 3.
- household_children - Number of children in household, top-coded to 3.

Methodology

Our target variable is *h1n1_vaccine* - whether respondent received H1N1 vaccine. This is a binary variable: 0 = No, respondent didn't get vaccine; 1 = Yes, respondent get vaccine.

Model performance will be evaluated according to the area under the receiver operating characteristic curve (AUC) for the target variables *h1n1_vaccine*. A receiving operator characteristic curve (ROC) is a tool for assessing the predictive accuracy of a binary classification model. Area under the curve can be up to 100%. The higher AUC, the better model fits data. The competition uses AUC as its evaluation metric, so our outcome must be the probabilities that a person received H1N1 vaccine, not binary labels.

Measurement of ROC is based on the confusion matrix, whose x-axis is False Positive Rate (FPR), and y-axis is True-Positive Rate (TPR). *Sensitivity* of a classifier is how many of the TRUE ones got right, *specificity* is how many of the FALSE ones got right, and *Misclassification rate* just asks how many predictions got wrong.

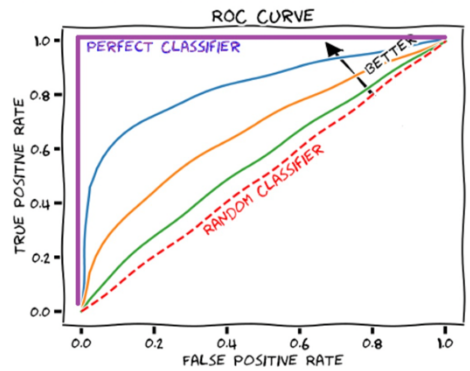
Confusion Matrix – Measure the type I and type II errors

True Negatives (TN) Predicted False Actual False	**False Positives (FP)** Predicted True Actual False
False Negatives (FN) Predicted False Actual True	**True Positives (TP)** Predicted True Actual True

$$Sensitivity = True\ Positive\ Rate = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$Specificity = True\ Negative\ Rate = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - \underbrace{\frac{FP}{N}}_{\text{False Positive Rate}}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum_{y_i \neq \hat{y}_i} 1 = \frac{(FP\ +\ FN)}{(N + P)} = Misclassification\ Rate$$



OLS Regression

The first model named *ols_full* contains all the features and adjusted R^2 is 0.3326. The second model is optimized after filtering out insignificant features and adjusted R^2 is 0.3299. The larger R-squared, the better regression model fits your data. It seems that optimization is not well effective because the R-squared is not quite different. However, the results confirmed some features have positive effects on receiving H1N1 vaccine. For instance, age > 55 years-old will more tends to get vaccination than the youths.

```
Call:
lm(formula = h1n1_vaccine ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.04503 -0.23520 -0.08938  0.21068  1.34183 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            -0.4063816  0.0674994  -6.021 1.80e-09 ***
h1n1_concern                            0.0032565  0.0051300   0.635 0.525583    
h1n1_knowledge                          0.0327621  0.0069211   4.734 2.24e-06 ***
behavioral_antiviral_meds               0.0053029  0.0178045   0.298 0.765829    
behavioral_avoidance                   -0.0074204  0.0097222  -0.763 0.445336    
behavioral_face_mask                    0.0063845  0.0159191   0.401 0.688385    
behavioral_wash_hands                  -0.0036640  0.0112445  -0.326 0.744549    
behavioral_large_gatherings            -0.0173149  0.0101737  -1.702 0.088803 .  
behavioral_outside_home                 0.0003631  0.0103854   0.035 0.972108    
behavioral_touch_face                   0.0064196  0.0094170   0.682 0.495446    
doctor_recc_h1n1                        0.3037252  0.0094170  32.253  < 2e-16 ***
chronic_med_condition                   0.0205963  0.0088502   2.327 0.019975 *  
child_under_6_months                    0.0397643  0.0140388   2.832 0.004629 ** 
health_worker                           0.1466924  0.0126316  11.613  < 2e-16 ***
health_insurance                        0.0689890  0.0128815   5.356 8.72e-08 ***
opinion_h1n1_vacc_effective             0.0783474  0.0041505  18.877  < 2e-16 ***
opinion_h1n1_risk                       0.0982310  0.0035213  27.896  < 2e-16 ***
opinion_h1n1_sick_from_vacc            -0.0053576  0.0031073  -1.724 0.084699 .  
age_group35 - 44 Years                 -0.0030571  0.0139486  -0.219 0.826525    
age_group45 - 54 Years                 -0.0048162  0.0126637  -0.380 0.703719    
age_group55 - 64 Years                  0.0601063  0.0130509   4.606 4.17e-06 ***
age_group65+ Years                      0.0806138  0.0141111   5.713 1.14e-08 ***
education< 12 Years                    -0.1044783  0.0499442  -2.092 0.036474 *  
education12 Years                      -0.0710011  0.0487569  -1.456 0.145363    
educationCollege Graduate              -0.0330952  0.0486490  -0.680 0.496341    
educationSome College                  -0.0582258  0.0486812  -1.196 0.231701    
raceHispanic                            0.0593123  0.0207074   2.864 0.004188 ** 
raceOther or Multiple                   0.0782352  0.0209497   3.734 0.000189 ***
raceWhite                               0.0711821  0.0148147   4.805 1.57e-06 ***
sexMale                                 0.0475699  0.0082346   5.777 7.85e-09 ***
income_poverty<= $75,000, Above Poverty -0.0375428  0.0128751  -2.916 0.003555 ** 
income_poverty> $75,000                -0.0210322  0.0147222  -1.429 0.153151    
income_povertyBelow Poverty            -0.0488599  0.0172274  -2.836 0.004575 ** 
employment_statusEmployed              -0.0513660  0.0462998  -1.109 0.267276    
employment_statusNot in Labor Force    -0.0455550  0.0464037  -0.982 0.326267    
employment_statusUnemployed            -0.0436792  0.0485291  -0.900 0.368110    
household_adults                       -0.0008568  0.0052718  -0.163 0.870898    
household_children                      0.0008366  0.0049505   0.169 0.865800    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3725 on 9515 degrees of freedom
Multiple R-squared:  0.3352,	Adjusted R-squared:  0.3326 
F-statistic: 129.6 on 37 and 9515 DF,  p-value: < 2.2e-16
```

Figure 1. Full Variables OLS Regression Results

```
Call:
lm(formula = h1n1_vaccine ~ h1n1_knowledge + doctor_recc_h1n1 + 
    child_under_6_months + health_worker + health_insurance + 
    opinion_h1n1_vacc_effective + opinion_h1n1_risk + age_group + 
    race + sex, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.07895 -0.23751 -0.08588  0.21171  1.32743 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -0.5849257  0.0238622 -24.513  < 2e-16 ***
h1n1_knowledge               0.0439498  0.0066257   6.633 3.46e-11 ***
doctor_recc_h1n1             0.3054702  0.0093314  32.736  < 2e-16 ***
child_under_6_months         0.0359599  0.0139931   2.570   0.0102 *  
health_worker                0.1508297  0.0123707  12.192  < 2e-16 ***
health_insurance             0.0828767  0.0123101   6.732 1.76e-11 ***
opinion_h1n1_vacc_effective  0.0787409  0.0040401  19.490  < 2e-16 ***
opinion_h1n1_risk            0.0958808  0.0032802  29.230  < 2e-16 ***
age_group35 - 44 Years       0.0029464  0.0135605   0.217   0.8280    
age_group45 - 54 Years       0.0002866  0.0123350   0.023   0.9815    
age_group55 - 64 Years       0.0654181  0.0121864   5.368 8.14e-08 ***
age_group65+ Years           0.0841752  0.0120478   6.987 3.00e-12 ***
raceHispanic                 0.0507913  0.0205238   2.475   0.0134 *  
raceOther or Multiple        0.0825627  0.0208678   3.956 7.66e-05 ***
raceWhite                    0.0795146  0.0144848   5.490 4.13e-08 ***
sexMale                      0.0506710  0.0079326   6.388 1.76e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3733 on 9537 degrees of freedom
Multiple R-squared:  0.3309,	Adjusted R-squared:  0.3299 
F-statistic: 314.5 on 15 and 9537 DF,  p-value: < 2.2e-16
```

Figure 2. Optimized OLS Regression Results

Logistic Regression

Logistic regression is used for modeling categorical outcomes, particularly no/yes, 0/1 outcomes. Such problems are called (binary) classification problems. We use *h1n1_vaccine* as our outcome metric and it fulfills the logistic regression model exactly.

```
Call:
glm(formula = h1n1_vaccine ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6118  -0.6292  -0.3853   0.5196   3.4694

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -6.603912   0.497876 -13.264  < 2e-16 ***
h1n1_concern                         0.047383   0.038365   1.235 0.216812
h1n1_knowledge                       0.223705   0.051164   4.372 1.23e-05 ***
behavioral_antiviral_meds            0.054936   0.128770   0.427 0.669654
behavioral_avoidance                -0.052921   0.072266  -0.732 0.463977
behavioral_face_mask                 0.055617   0.112189   0.496 0.620076
behavioral_wash_hands                0.008729   0.086464   0.101 0.919589
behavioral_large_gatherings         -0.132883   0.074359  -1.787 0.073930 .
behavioral_outside_home             -0.011669   0.075668  -0.154 0.877445
behavioral_touch_face                0.034178   0.069274   0.493 0.621754
doctor_recc_h1n1                     1.638919   0.061523  26.639  < 2e-16 ***
chronic_med_condition                0.131787   0.063404   2.079 0.037661 *
child_under_6_months                 0.261787   0.101389   2.582 0.009823 **
health_worker                        0.896370   0.086426  10.372  < 2e-16 ***
health_insurance                     0.679779   0.109768   6.193 5.91e-10 ***
opinion_h1n1_vacc_effective          0.715664   0.037826  18.920  < 2e-16 ***
opinion_h1n1_risk                    0.571238   0.024547  23.271  < 2e-16 ***
opinion_h1n1_sick_from_vacc         -0.021035   0.023164  -0.908 0.363831
age_group35 - 44 Years              -0.067294   0.105293  -0.639 0.522748
age_group45 - 54 Years              -0.093011   0.097244  -0.956 0.338831
age_group55 - 64 Years               0.373615   0.096551   3.870 0.000109 ***
age_group65+ Years                   0.509887   0.104186   4.894 9.88e-07 ***
education< 12 Years                 -0.914986   0.350470  -2.611 0.009035 **
education12 Years                   -0.636020   0.339101  -1.876 0.060710 .
educationCollege Graduate           -0.386693   0.337919  -1.144 0.252483
educationSome College               -0.557993   0.338334  -1.649 0.099099 .
raceHispanic                         0.565242   0.163073   3.466 0.000528 ***
raceOther or Multiple                0.706712   0.161840   4.367 1.26e-05 ***
raceWhite                            0.616345   0.121302   5.081 3.75e-07 ***
sexMale                              0.346906   0.060714   5.714 1.10e-08 ***
income_poverty<= $75,000, Above Poverty -0.303538   0.096446  -3.147 0.001648 **
income_poverty> $75,000             -0.196620   0.108964  -1.804 0.071162 .
income_povertyBelow Poverty         -0.423226   0.132401  -3.197 0.001391 **
employment_statusEmployed           -0.320311   0.345961  -0.926 0.354520
employment_statusNot in Labor Force -0.265648   0.346048  -0.768 0.442688
employment_statusUnemployed         -0.313165   0.365357  -0.857 0.391364
household_adults                    -0.014680   0.039317  -0.373 0.708873
household_children                   0.011766   0.037251   0.316 0.752106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3. Logistic Regression Results

	Sensitivity (TPR)	Specificity (TNR)	Misclassification Rate	AUC
Logistic Regression	0.56	0.921	0.188	0.8535
LDA	0.57	0.917	0.189	0.8534
QDA	0.67	0.826	0.221	0.8127

Table 1. Comparison of TPR, TNR, MR and AUC

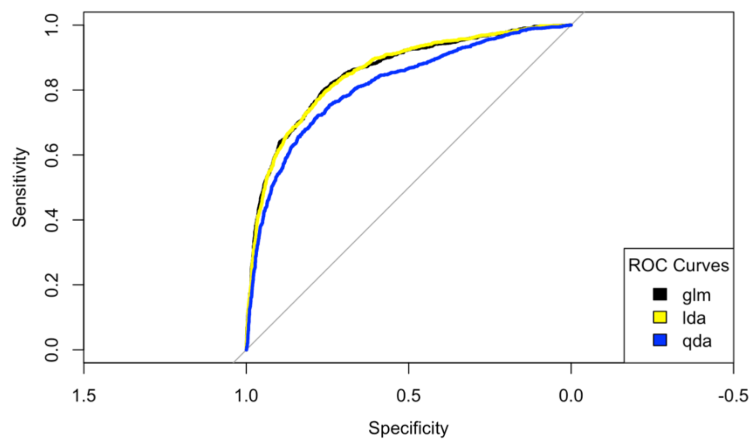


Figure 4. ROC comparison of LDA, QDA and Logistic Regression

LDA & QDA

To illustrate comparably we import Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Same as in logistic regression, LDA affine decision boundaries particularly. QDA has far more parameters to be estimated.

Classification and Regression Tree (CART)

Trees successively split the features into two areas (greater and smaller than a certain cutoff) and the cutoff level are chosen based on a *greedy approach*: we are seeking the split that will reduce the prediction error (sum-of-squared error for regression problems, different possible choices for classification problems).

The general approach is to build a large tree and then "prune" off nodes (or subtrees). The approach here is to find the nodes that, again, lead to a minimal increase in the prediction error. This yields a sequence of trees of ordered size complexity and determining a suitable predictive model will require to trade off simplicity (low variance) and complexity (low bias). This can be accomplished via cross validation. So that the final tree can be thought of as a model with constant predictions in square partitions of the features.

According to cross-validation table, we get the minimal prediction error of 0.64572 when tree split to 7 nodes seen as below.

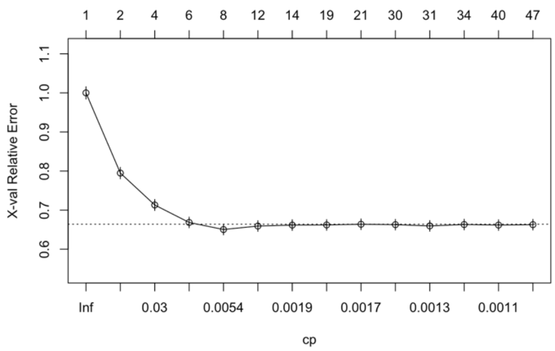


Figure 5. cross-validation table

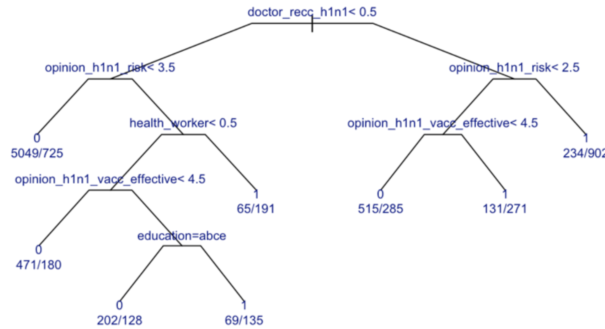


Figure 6. Pruned tree with # nodes = 7

Random Forest

Random forests arrive at their predictions by fitting trees to bootstrap replications of the dataset, just as in bagging. However, they additionally sample from the set of features to reduce the correlation of the predictors, so to take advantage of the diversification benefit.

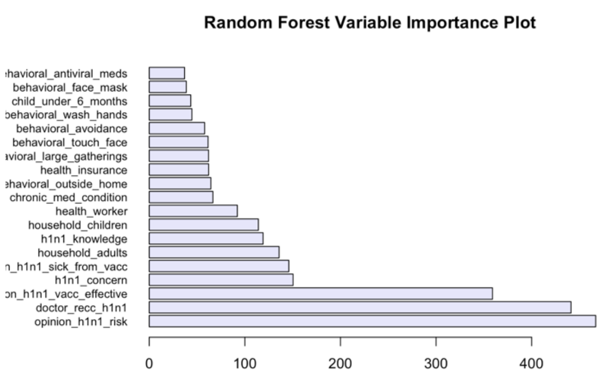


Figure 7. Random Forest Variable Importance Plot

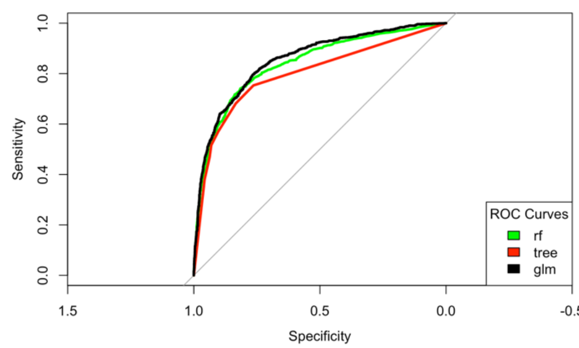


Figure 8. ROC comparison of RF, CART and Logistic Regression

Artificial Neural Nets

A neural net generally consists of an input layer with the features $\{X_1, \dots, X_p\}$, one or more hidden layers with neurons $\{Z_1, \dots, Z_M\}$, and an output layer Y. In the case of a one-dimensional regression problem, the output layer consists of a single outcome Y. In a single-layer neural network, the inputs are processed into the neurons of the hidden layer, which in turn are processed into the outputs. More precisely, for each neuron Z_M , the X_p's are linearly aggregated and transformed via a sigmoid function:

$$\boxed{\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_{m,1} X_1 + \alpha_{m,2} X_2 + \ldots + \alpha_{m,p} X_p) \\ Y &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_M Z_M + \varepsilon \end{aligned}}$$

The sigmoid function has the appealing property that it can depict highly linear and highly non-linear relationships. The constant term $(\alpha_{0,m})$ together with the norm of the coefficient vector $(\alpha_{m,1}, \dots, \alpha_{m,p})$ determines how nonlinear the relationship in that neuron is. The neurons are then aggregated to the response Y. In a deep neural network, there are several hidden layers in which the neurons are processed into other neurons.

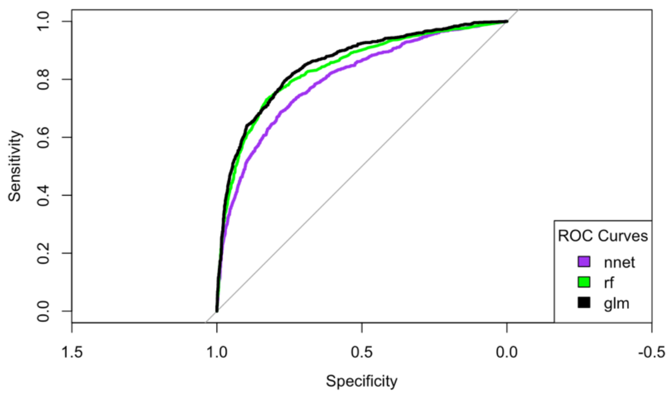


Figure 9. ROC comparison of NN, RF and Logistic Regression

	Sensitivity (TPR)	Specificity (TNR)	Misclassification Rate	AUC
CART	0.52	0.93	0.197	0.7979
Random Forest	0.54	0.92	0.194	0.8386
Neural Nets	1	0.0043	0.690	0.7994

Table 2. Comparison of TPR, TNR, MR and AUC

Results

Under the consideration of AUC, obviously the highest score is .8535 by optimized logistic regression and the lowest score is 0.7979 by classification and random trees, which means logistic regression model is the best model on fitting this dataset.

Also, we can learn 3 things from the regression results. First, elder people more tend on get H1N1 vaccination. Second, people who hold a higher-level knowledge and realize more on the risk of H1N1 flu will more tend to get vaccination. Third, healthy workers with medical insurance will more tends to get vaccination.

Conclusion

What kind of features will influence people to receive H1N1 vaccine? How they affect the probability of vaccination reception? Those features may come from age, sex, race, personality, routine behavior, others' opinion, etc. And we conduct this analysis in order to find the best fitted model, trying to answer those two questions as perfect as possible.

Who can benefit from this analysis? I suppose the answer will be government agencies (i.e., CDC), medical companies, insurance companies, WHO, and some social non-profit organizations. For instance, once we know the level of knowledge on H1N1 flu can heavily affect the vaccination reception probability. CDC may act on increasing the vaccination rate by spreading knowledge related with H1N1 in public.

References

Bevans, R. (2020, March 26). An introduction to the Akaike information criterion. Retrieved from Scribbr: https://www.scribbr.com/statistics/akaike-information-criterion/

DrivenData. (2021). Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines. Retrieved from DRIVENDATA: https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/

Fernando, J. (2021, September 12). R-Squared. Retrieved from Investopedia: https://www.investopedia.com/terms/r/r-squared.asp#:~:text=of%20R%2DSquared-,What%20Is%20R%2DSquared%3F,variables%20in%20a%20regression%20model.