Random Forest Workflow

Vanilton Paulo and Nhi Nguyen

2025-08-13

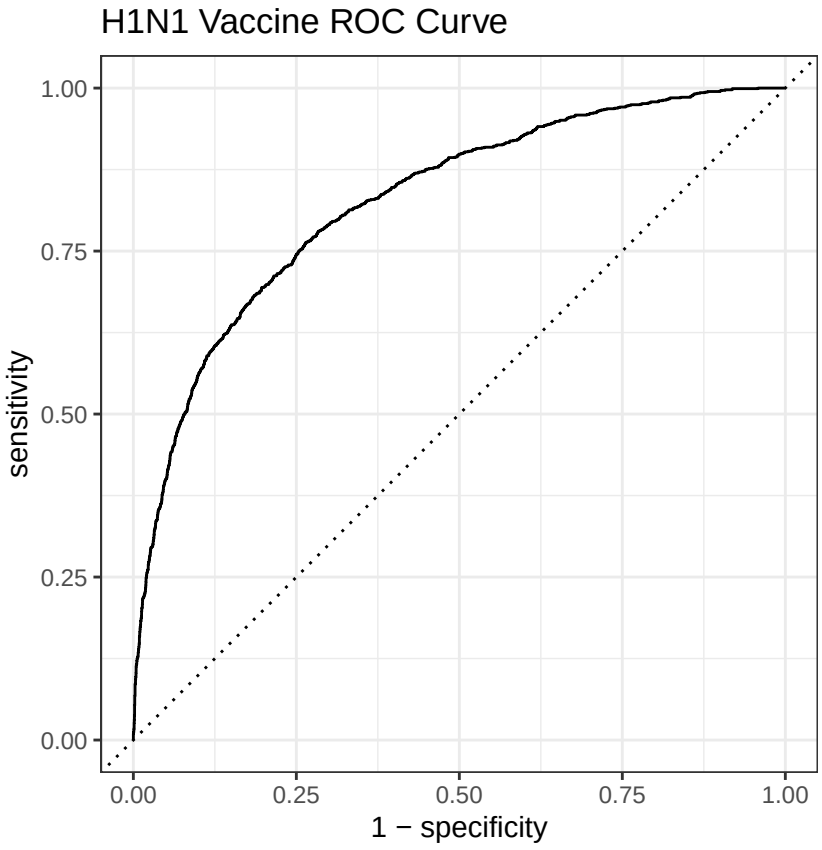
Data Overview

```
## Rows: 26,707
## Columns: 38
## $ respondent_id               <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ~
## $ h1n1_concern                <dbl> 1, 3, 1, 1, 2, 3, 0, 1, 0, 2, 2, 1, 1, 1, ~
## $ h1n1_knowledge              <dbl> 0, 2, 1, 1, 1, 1, 0, 0, 2, 1, 1, 2, 1, 1, ~
## $ behavioral_antiviral_meds   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ behavioral_avoidance        <dbl> 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, ~
## $ behavioral_face_mask        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ behavioral_wash_hands       <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, ~
## $ behavioral_large_gatherings <dbl> 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, ~
## $ behavioral_outside_home     <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ~
## $ behavioral_touch_face       <dbl> 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, ~
## $ doctor_recc_h1n1            <dbl> 0, 0, NA, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,~
## $ doctor_recc_seasonal        <dbl> 0, 0, NA, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ chronic_med_condition       <dbl> 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, ~
## $ child_under_6_months        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, ~
## $ health_worker               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ health_insurance            <dbl> 1, 1, NA, NA, NA, NA, NA, 1, NA, 1, 0, 1, ~
## $ opinion_h1n1_vacc_effective <dbl> 3, 5, 3, 3, 3, 5, 4, 5, 4, 4, 4, 3, 3, 3, ~
## $ opinion_h1n1_risk           <dbl> 1, 4, 1, 3, 3, 2, 1, 2, 1, 2, 1, 2, 2, 1, ~
## $ opinion_h1n1_sick_from_vacc <dbl> 2, 4, 1, 5, 2, 1, 1, 1, 1, 2, 2, 2, 1, 4, ~
## $ opinion_seas_vacc_effective <dbl> 2, 4, 4, 5, 3, 5, 4, 4, 4, 4, 5, 4, 5, 4, ~
## $ opinion_seas_risk           <dbl> 1, 2, 1, 4, 1, 4, 2, 2, 2, 2, 4, 2, 4, 2, ~
## $ opinion_seas_sick_from_vacc <dbl> 2, 4, 2, 1, 4, 4, 1, 1, 1, 2, 4, 1, 1, 4, ~
## $ age_group                   <chr> "55 - 64 Years", "35 - 44 Years", "18 - 34~
## $ education                   <chr> "< 12 Years", "12 Years", "College Graduat~
## $ race                        <chr> "White", "White", "White", "White", "White~
## $ sex                         <chr> "Female", "Male", "Male", "Female", "Femal~
## $ income_poverty              <chr> "Below Poverty", "Below Poverty", "<= $75,~
## $ marital_status              <chr> "Not Married", "Not Married", "Not Married~
## $ rent_or_own                 <chr> "Own", "Rent", "Own", "Rent", "Own", "Own"~
## $ employment_status           <chr> "Not in Labor Force", "Employed", "Employe~
## $ hhs_geo_region              <chr> "oxchjgsf", "bhuqouqj", "qufhixun", "lrirc~
## $ census_msa                  <chr> "Non-MSA", "MSA, Not Principle  City", "MS~
## $ household_adults            <dbl> 0, 0, 2, 0, 1, 2, 0, 2, 1, 0, 2, 1, 1, 1, ~
## $ household_children          <dbl> 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 2, 0, 2, ~
## $ employment_industry         <chr> NA, "pxcmvdjn", "rucpziij", NA, "wxleyezf"~
## $ employment_occupation       <chr> NA, "xgwztkwe", "xtkaffoo", NA, "emcorrxb"~
## $ h1n1_vaccine                <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, ~
## $ seasonal_vaccine            <dbl> 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, ~
```

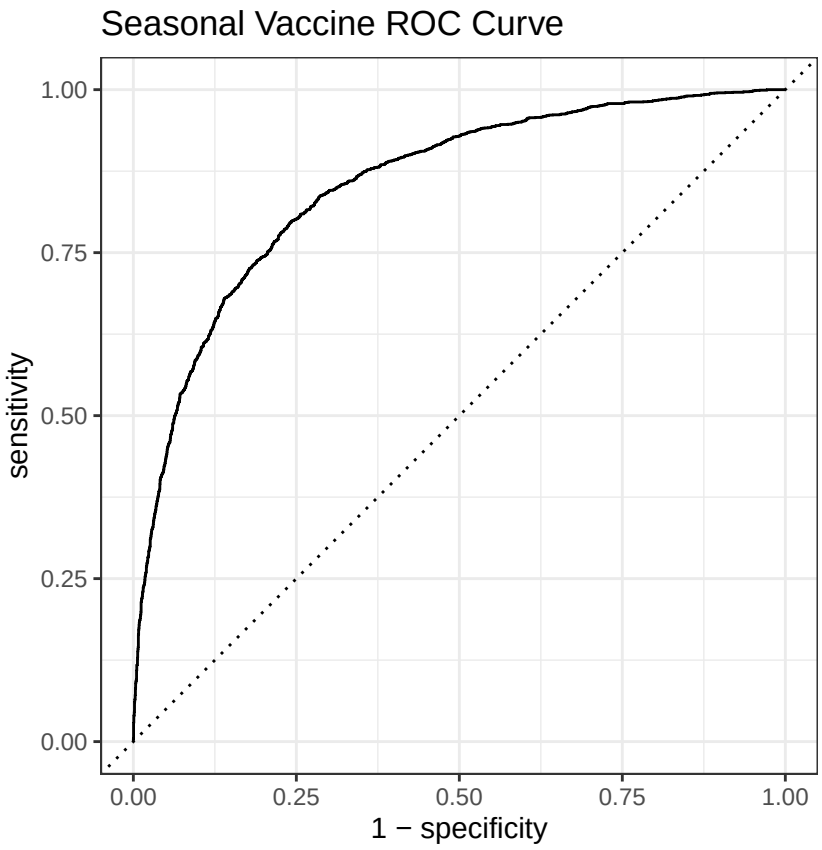
Pre-tuning

Plot H1N1 Vaccine and Seasonal Vaccine ROC Curve

H1N1 Vaccine



Seasonal Vaccine



Calcualte the ROC AUC for H1N1 and Seasonal Vaccine

H1N1 Vaccine

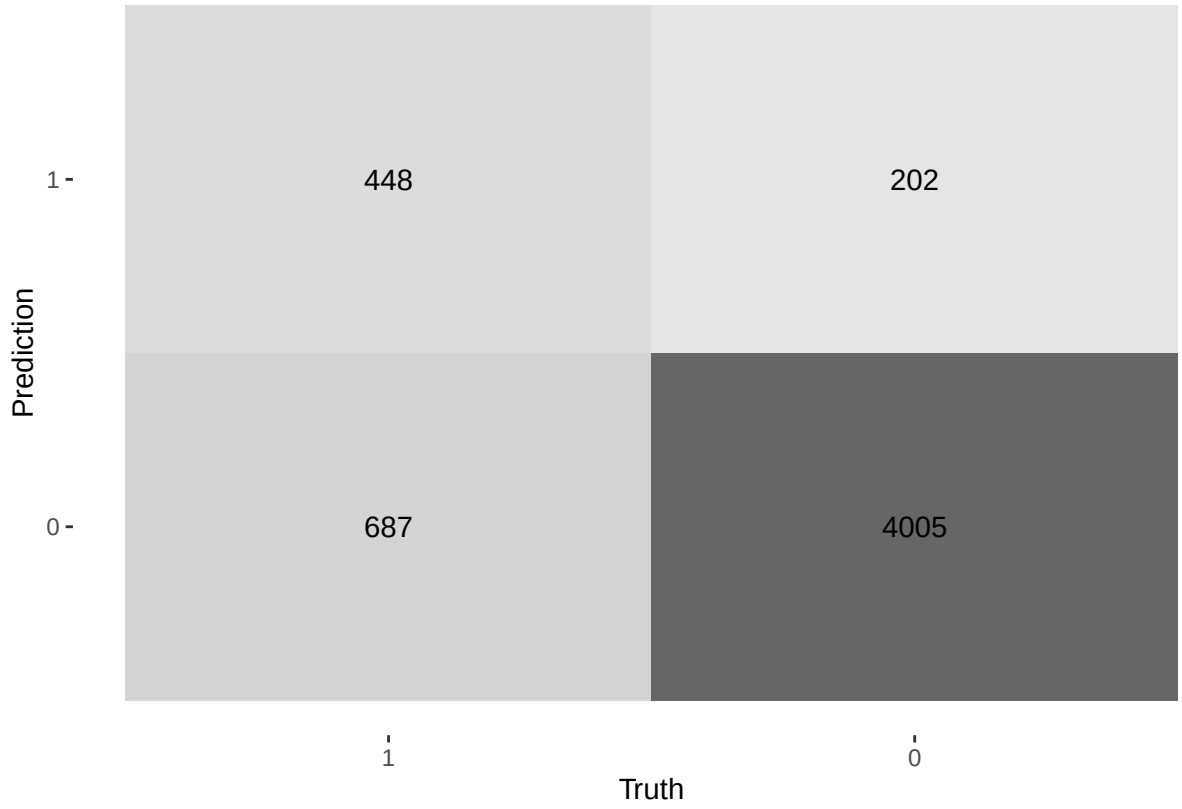
```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.826
```

Seasonal Vaccine

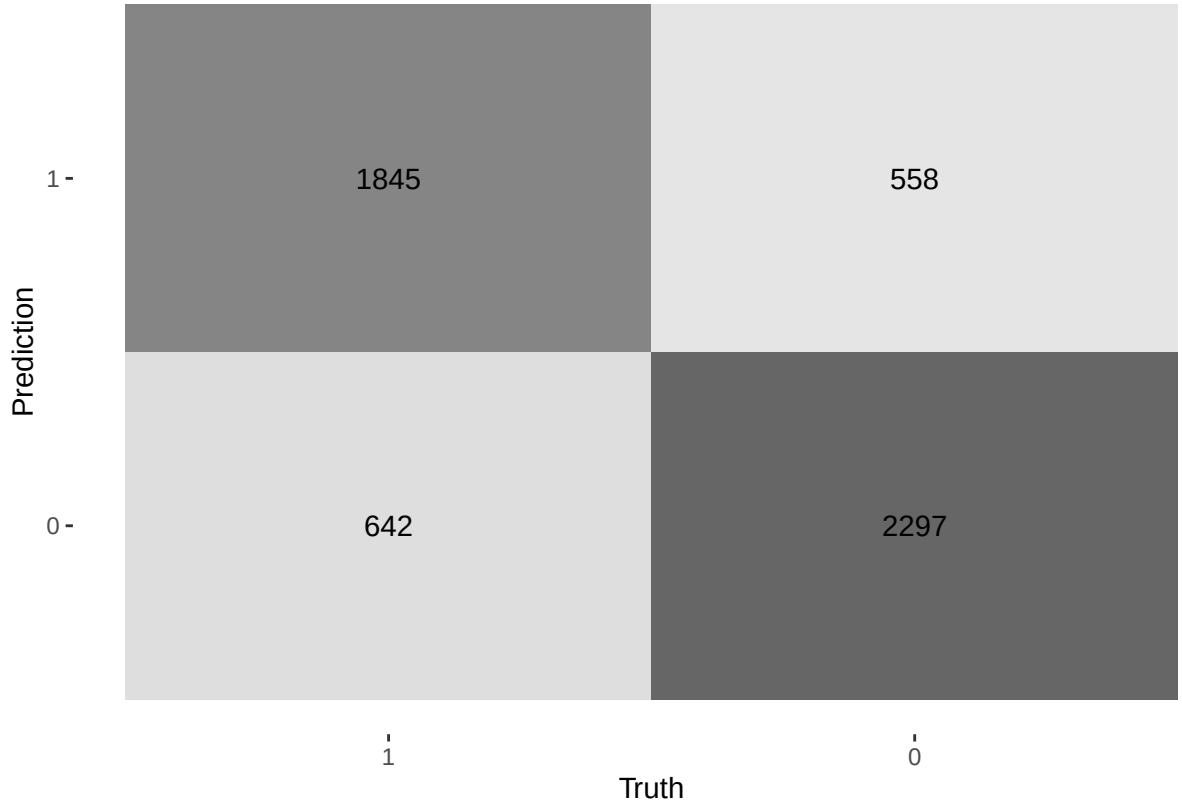
```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.853
```

Confusion matrix with Heatmap for H1N1 and Seasonal Vaccine

H1N1 Vaccine

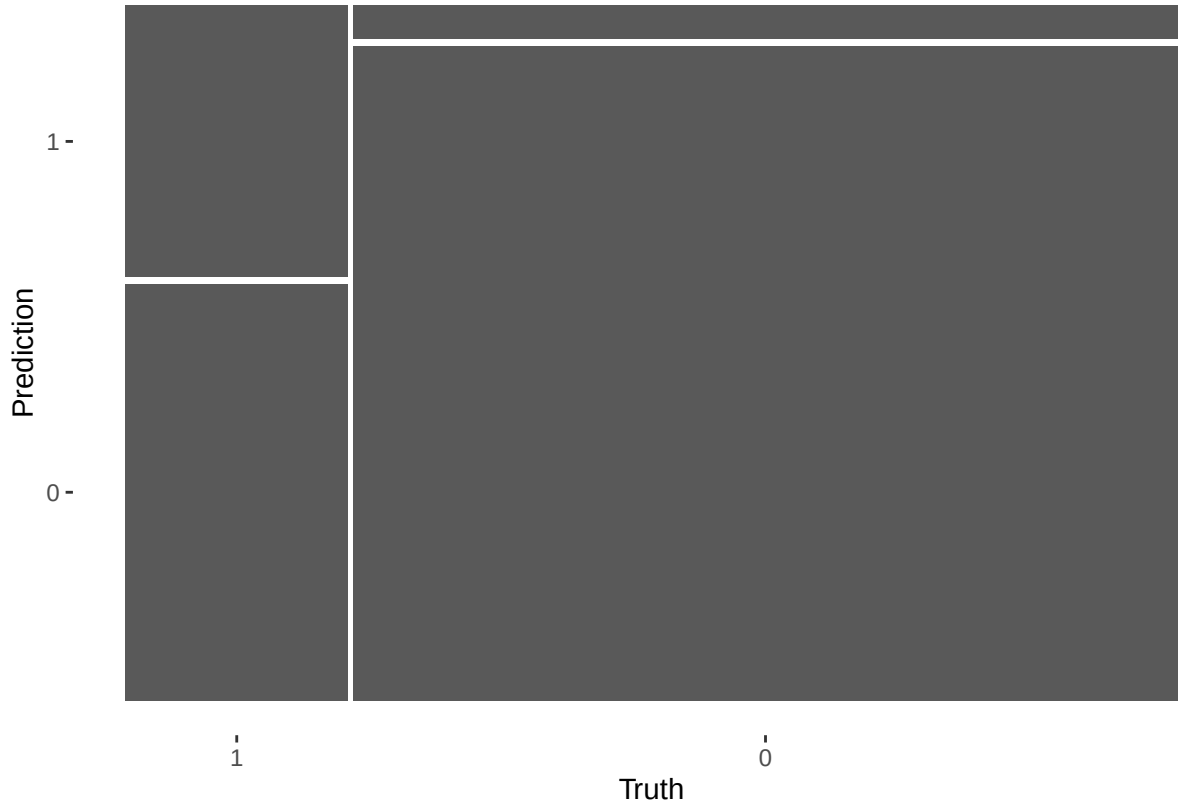


Seasonal Vaccine

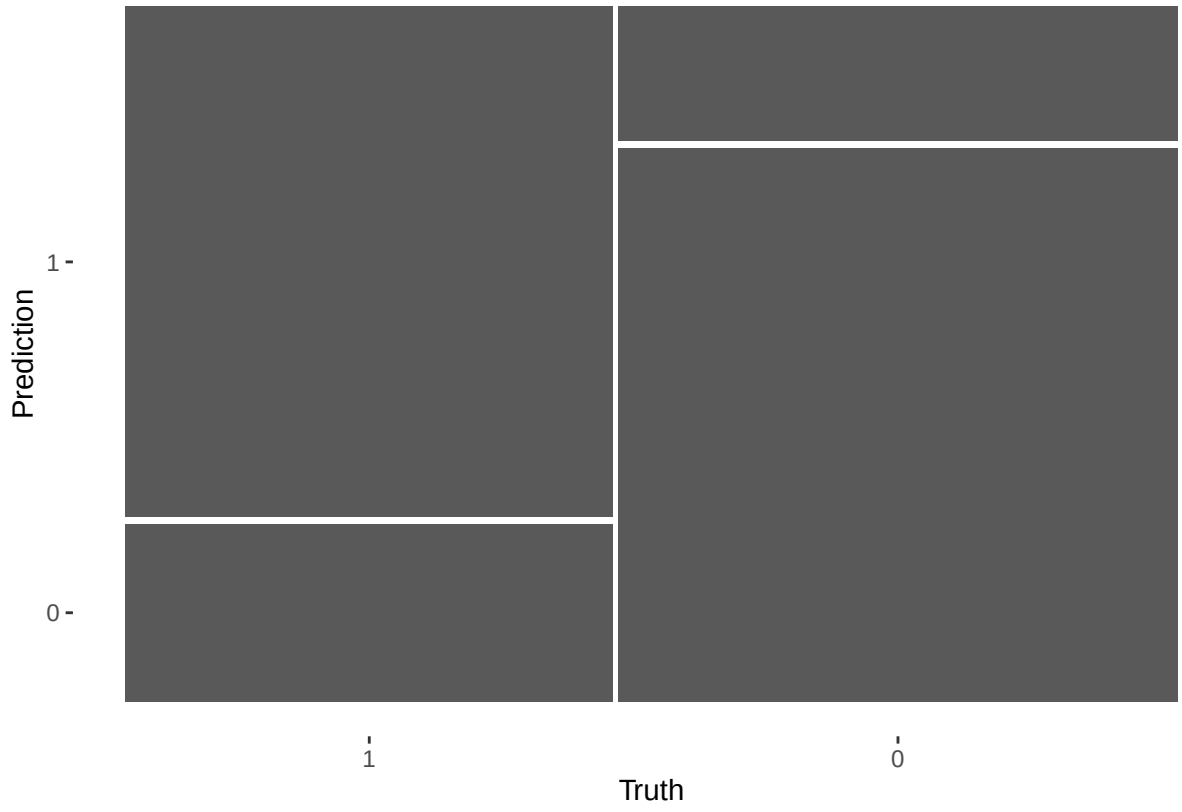


Compute confusion matrix with mosaic plots for H1N1 and Seasonal Vaccine

H1N1 Vaccine



Seasonal Vaccine



Custom metric predictions for H1N1 and Seasonal Vaccine

H1N1 Vaccine

```
## # A tibble: 5 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.834
## 2 sens     binary         0.395
## 3 spec     binary         0.952
## 4 f_meas   binary         0.502
## 5 roc_auc  binary         0.826
```

Seasonal Vaccine

```
## # A tibble: 5 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.775
## 2 sens     binary         0.742
## 3 spec     binary         0.805
## 4 f_meas   binary         0.755
## 5 roc_auc  binary         0.853
```

Cross Validation

Detailed cross validation results for H1N1 and Seasonal Vaccine

H1N1 Vaccine

```
## # A tibble: 4 x 5
##   .metric    min median   max      sd
##   <chr>    <dbl>  <dbl> <dbl>   <dbl>
## 1 accuracy 0.826  0.840 0.848 0.00740
## 2 roc_auc  0.818  0.833 0.850 0.0112
## 3 sens     0.360  0.405 0.449 0.0281
## 4 spec     0.952  0.956 0.962 0.00367
```

Seasonal Vaccine

```
## # A tibble: 4 x 5
##   .metric    min median   max      sd
##   <chr>    <dbl>  <dbl> <dbl>   <dbl>
## 1 accuracy 0.758  0.785 0.793 0.00985
## 2 roc_auc  0.832  0.857 0.872 0.0102
## 3 sens     0.732  0.753 0.769 0.0100
## 4 spec     0.768  0.809 0.833 0.0185
```

Hyperparameter tuning

View results for H1N1 and Seasonal Vaccine

H1N1 Vaccine

```
## # A tibble: 2,000 x 10
##     mtry trees min_n sample.fraction .metric  .estimator   mean     n  std_err
##    <int> <int> <int>           <dbl> <chr>    <chr>       <dbl> <int>    <dbl>
##  1     0   932    14           0.805 accuracy binary     0.838     10 0.00234
##  2     0   932    14           0.805 roc_auc  binary     0.835     10 0.00349
##  3     0   932    14           0.805 sens     binary     0.400     10 0.00930
##  4     0   932    14           0.805 spec     binary     0.956     10 0.00108
##  5     1  1962    15           0.309 accuracy binary     0.791     10 0.000408
##  6     1  1962    15           0.309 roc_auc  binary     0.821     10 0.00402
##  7     1  1962    15           0.309 sens     binary     0.0174    10 0.00193
##  8     1  1962    15           0.309 spec     binary     0.999     10 0.000217
##  9     0  1390    32           0.796 accuracy binary     0.838     10 0.00215
## 10     0  1390    32           0.796 roc_auc  binary     0.836     10 0.00352
## # i 1,990 more rows
## # i 1 more variable: .config <chr>
```

Explore detailed ROC AUC results for each fold

```
## # A tibble: 10 x 4
##    id     min_roc_auc median_roc_auc max_roc_auc
##    <chr>        <dbl>          <dbl>       <dbl>
##  1 Fold01       0.766          0.814       0.822
```

```
##  2 Fold02       0.771          0.825       0.833
##  3 Fold03       0.793          0.845       0.853
##  4 Fold04       0.778          0.834       0.843
##  5 Fold05       0.774          0.840       0.849
##  6 Fold06       0.779          0.824       0.832
##  7 Fold07       0.807          0.839       0.848
##  8 Fold08       0.792          0.835       0.842
##  9 Fold09       0.762          0.824       0.832
## 10 Fold10       0.762          0.813       0.822
```

Seasonal Vaccine

```
## # A tibble: 2,000 x 10
##     mtry trees min_n sample.fraction .metric  .estimator  mean     n std_err
##    <int> <int> <int>           <dbl> <chr>    <chr>      <dbl> <int>   <dbl>
##  1     1  1580    19           0.885 accuracy binary     0.765    10 0.00366
##  2     1  1580    19           0.885 roc_auc  binary     0.842    10 0.00333
##  3     1  1580    19           0.885 sens     binary     0.676    10 0.00367
##  4     1  1580    19           0.885 spec     binary     0.842    10 0.00509
##  5     0   117    32           0.370 accuracy binary     0.781    10 0.00318
##  6     0   117    32           0.370 roc_auc  binary     0.855    10 0.00295
##  7     0   117    32           0.370 sens     binary     0.748    10 0.00301
##  8     0   117    32           0.370 spec     binary     0.810    10 0.00541
##  9     0  1569    29           0.201 accuracy binary     0.782    10 0.00310
## 10     0  1569    29           0.201 roc_auc  binary     0.855    10 0.00309
## # i 1,990 more rows
## # i 1 more variable: .config <chr>
```

Explore detailed ROC AUC results for each fold

```
## # A tibble: 10 x 4
##    id     min_roc_auc median_roc_auc max_roc_auc
##    <chr>        <dbl>          <dbl>       <dbl>
##  1 Fold01       0.754          0.849       0.860
##  2 Fold02       0.748          0.845       0.861
##  3 Fold03       0.712          0.817       0.834
##  4 Fold04       0.736          0.848       0.860
##  5 Fold05       0.756          0.859       0.873
##  6 Fold06       0.756          0.843       0.860
##  7 Fold07       0.756          0.848       0.858
##  8 Fold08       0.753          0.842       0.854
##  9 Fold09       0.746          0.841       0.856
## 10 Fold10       0.738          0.843       0.857
```

Post-tunning

Selecting the best model for H1N1 and Seasonal Vaccine

```
## [1] "H1N1 Vaccine"
```

```
## # A tibble: 5 x 10
```

```
##    mtry trees min_n sample.fraction .metric .estimator  mean     n std_err
##   <int> <int> <int>           <dbl> <chr>   <chr>      <dbl> <int>   <dbl>
## 1     0   857    40           0.843 roc_auc binary     0.837    10 0.00348
## 2     0  1213    40           0.890 roc_auc binary     0.837    10 0.00356
## 3     0  1702    39           0.989 roc_auc binary     0.837    10 0.00360
## 4     0  1867    36           0.885 roc_auc binary     0.837    10 0.00359
## 5     0  1562    32           0.847 roc_auc binary     0.837    10 0.00357
## # i 1 more variable: .config <chr>
```

```
## [1] "Seasonal Vaccine"
```

```
## # A tibble: 5 x 10
##    mtry trees min_n sample.fraction .metric .estimator  mean     n std_err
##   <int> <int> <int>           <dbl> <chr>   <chr>      <dbl> <int>   <dbl>
## 1     0  1700    37           0.934 roc_auc binary     0.857    10 0.00310
## 2     0  1904    34           0.984 roc_auc binary     0.857    10 0.00314
## 3     0  1375    35           0.922 roc_auc binary     0.857    10 0.00311
## 4     0  1984    32           0.956 roc_auc binary     0.857    10 0.00312
## 5     0  1948    23           0.906 roc_auc binary     0.856    10 0.00311
## # i 1 more variable: .config <chr>
```

H1N1 Vaccine -> Tuned parameters

```
## # A tibble: 1 x 5
##    mtry trees min_n sample.fraction .config
##   <int> <int> <int>           <dbl> <chr>
## 1     0   857    40           0.843 Preprocessor1_Model387
```

Seasonal Vaccine -> Tuned parameters

```
## # A tibble: 1 x 5
##    mtry trees min_n sample.fraction .config
##   <int> <int> <int>           <dbl> <chr>
## 1     0  1700    37           0.934 Preprocessor1_Model105
```

Metrics of post tuned model

```
## [1] "H1N1 Vaccine"
```

```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.833 Preprocessor1_Model1
## 2 roc_auc     binary         0.828 Preprocessor1_Model1
## 3 brier_class binary         0.121 Preprocessor1_Model1
```

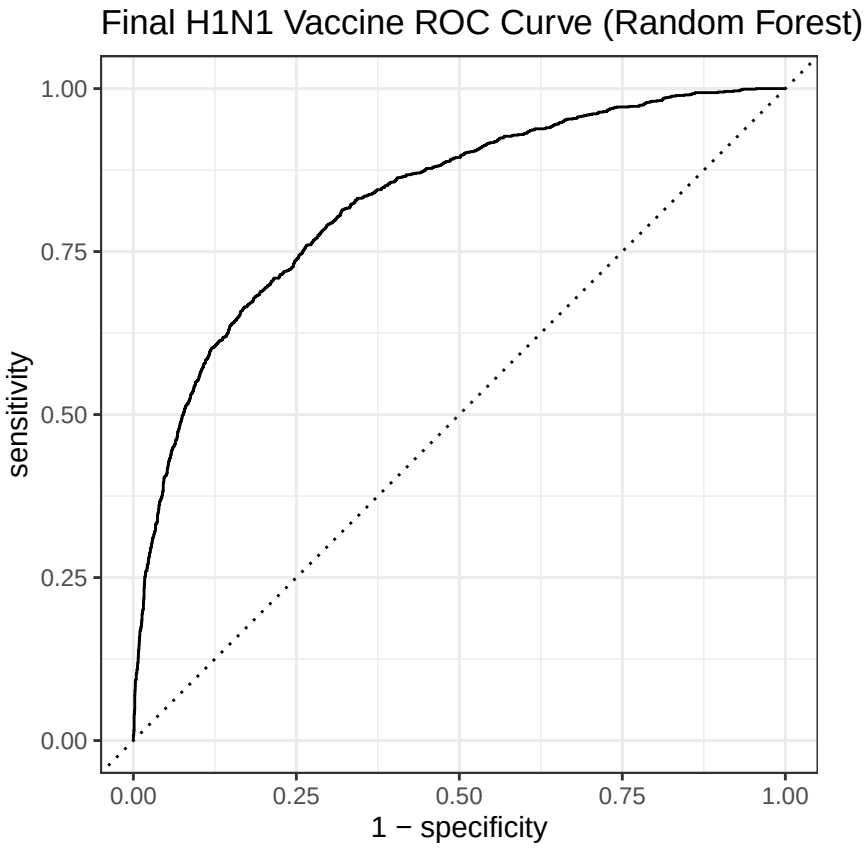
```
## [1] "Seasonal Vaccine"
```

```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
```

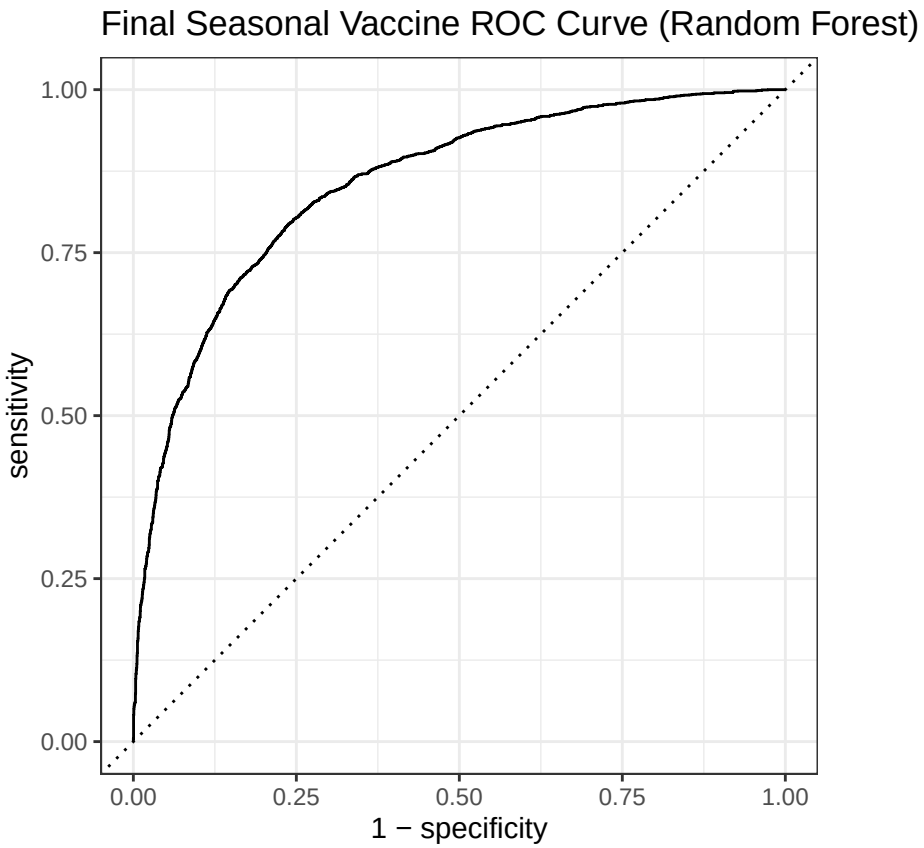
```
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.774 Preprocessor1_Model1
## 2 roc_auc     binary         0.854 Preprocessor1_Model1
## 3 brier_class binary         0.156 Preprocessor1_Model1
```

Roc Curve of post tuned model for H1N1 and Seasonal Vaccine

H1N1 Vaccine

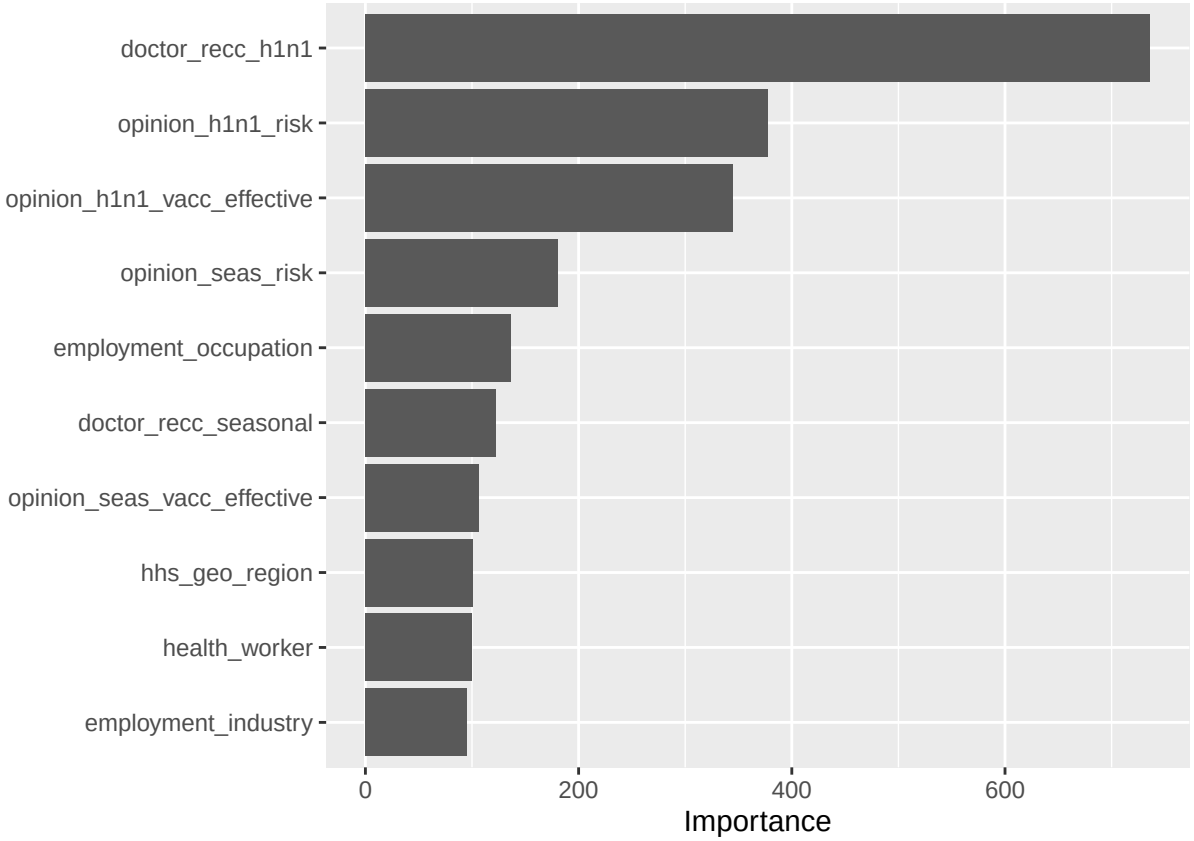


Seasonal Vaccine

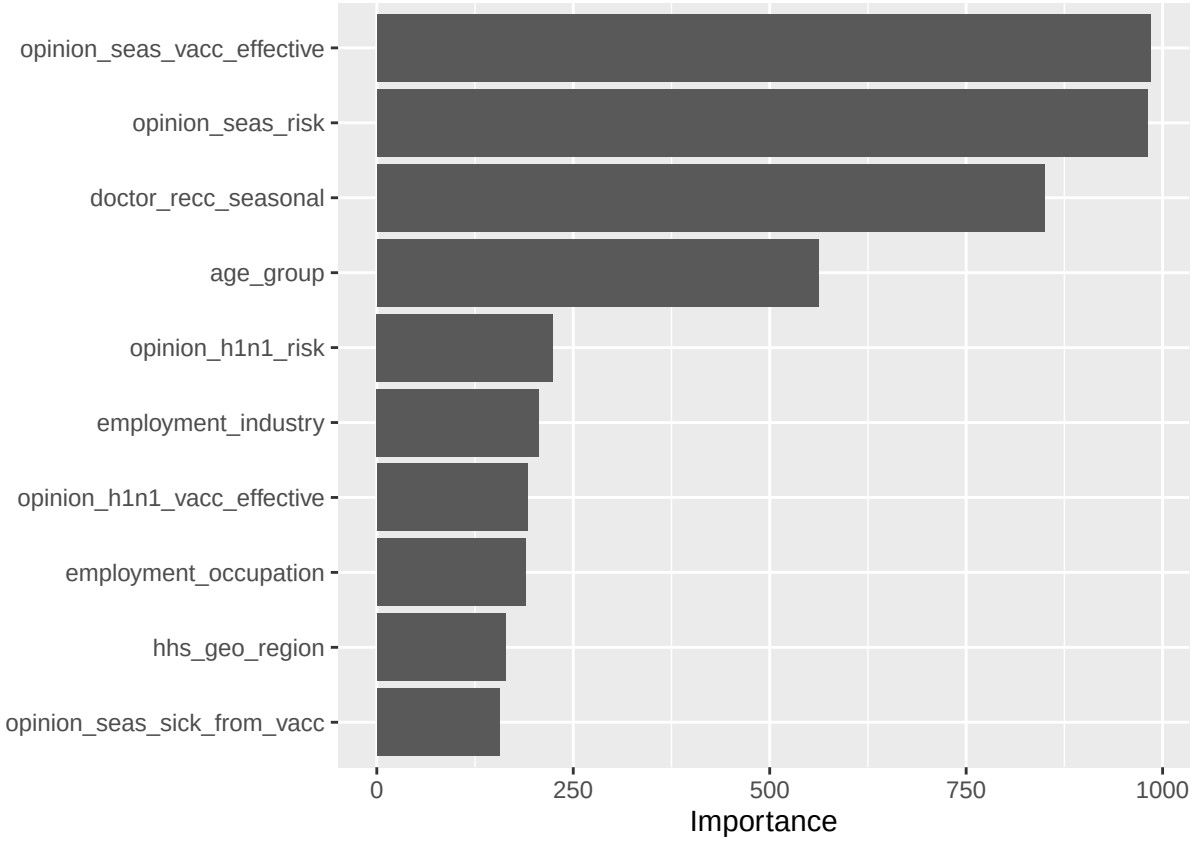


Variable importance for H1N1 and Seasonal Vaccine

H1N1 Vaccine



Seasonal Vaccine



Predictions on test data

```
## [1] "H1N1 Vaccine"
```

```
## [1] 0.14480251 0.06745837 0.42473046 0.56929261 0.26901016 0.56269021
```

```
## [1] "Seasonal Vaccine"
```

```
## [1] 0.30389822 0.06924356 0.80258354 0.88453575 0.45324796 0.90664767
```