Imperial College London

4 May 2021

Predicting seasonal flu vaccine uptake

Machine Learning Individual Project 2020-21

Camilla Callierotti

Imperial College London

Outline

Imperial College London

Outline

Imperial College London

Dataset

```
Dimensions of X: (26707, 36)
Dimensions of Y: (26707, 3)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26707 entries, 0 to 26706
Data columns (total 36 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   respondent_id                26707 non-null  int64
 1   h1n1_concern                 26615 non-null  float64
 2   h1n1_knowledge               26591 non-null  float64
 3   behavioral_antiviral_meds    26636 non-null  float64
 4   behavioral_avoidance         26499 non-null  float64
 5   behavioral_face_mask         26688 non-null  float64
 6   behavioral_wash_hands        26665 non-null  float64
 7   behavioral_large_gatherings  26620 non-null  float64
 8   behavioral_outside_home      26625 non-null  float64
 9   behavioral_touch_face        26579 non-null  float64
 10  doctor_recc_h1n1             24547 non-null  float64
 11  doctor_recc_seasonal         24547 non-null  float64
 12  chronic_med_condition        25736 non-null  float64
 13  child_under_6_months         25887 non-null  float64
 14  health_worker                25903 non-null  float64
 15  health_insurance             14433 non-null  float64
 16  opinion_h1n1_vacc_effective  26316 non-null  float64
 17  opinion_h1n1_risk            26319 non-null  float64
 18  opinion_h1n1_sick_from_vacc  26312 non-null  float64
 19  opinion_seas_vacc_effective  26245 non-null  float64
 20  opinion_seas_risk            26193 non-null  float64
 21  opinion_seas_sick_from_vacc  26170 non-null  float64
 22  age_group                    26707 non-null  object
 23  education                    25300 non-null  object
 24  race                         26707 non-null  object
 25  sex                          26707 non-null  object
 26  income_poverty               22284 non-null  object
 27  marital_status               25299 non-null  object
 28  rent_or_own                  24665 non-null  object
 29  employment_status            25244 non-null  object
 30  hhs_geo_region               26707 non-null  object
 31  census_msa                   26707 non-null  object
 32  household_adults             26458 non-null  float64
 33  household_children           26458 non-null  float64
 34  employment_industry          13377 non-null  object
 35  employment_occupation        13237 non-null  object
dtypes: float64(23), int64(1), object(12)
```

Figure 1: Original dataset features and structure

- 26'707 observations
- 36 features
  - Vaccine opinions
  - Behaviours
  - Sociodemographic factors

⇒ Dataset also contained information on H1N1 vaccine but the focus of this project is seasonal vaccine

Imperial College London

Outline

Missingness

- $>$50% observations missing: drop feature
- $<$50% observations missing: imputation
  - Mode imputation (categorical features)
  - Mean imputation (numerical features)
- Column containing indecipherable observations: drop feature

Imperial College London

Categorical features

- Label encoding (hierarchical features)
- One-hot encoding (non-hierarchical features)

	age_group	education	race	sex	income_poverty	marital_status	rent_or_own	employment_status	census_msa
0	55 - 64 Years	< 12 Years	White	Female	Below Poverty	Not Married	Own	Not in Labor Force	Non-MSA
1	35 - 44 Years	12 Years	White	Male	Below Poverty	Not Married	Rent	Employed	MSA, Not Principle City
2	18 - 34 Years	College Graduate	White	Male	<= $75,000, Above Poverty	Not Married	Own	Employed	MSA, Not Principle City
3	65+ Years	12 Years	White	Female	Below Poverty	Not Married	Rent	Not in Labor Force	MSA, Principle City
4	45 - 54 Years	Some College	White	Female	<= $75,000, Above Poverty	Married	Own	Employed	MSA, Not Principle City

Collinear features (1/2)

Feature engineering to merge collinear features into one normally distributed feature:

- Cleanliness: Antiviral meds, avoidance, face mask, hand washing, large gatherings, outside home, face touching
- Opinion on seasonal flu vaccine: Vaccine effectiveness, seasonal flu risk, getting seasonal flu from vaccine

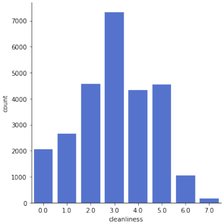

Figure 2: Cleanliness variable distribution

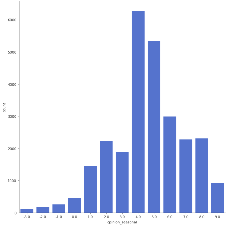

Figure 3: Opinion variable distribution

Imperial College London

Collinear features (2/2)

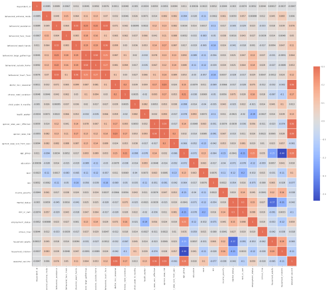

Figure 4: Heatmap before feature engineering

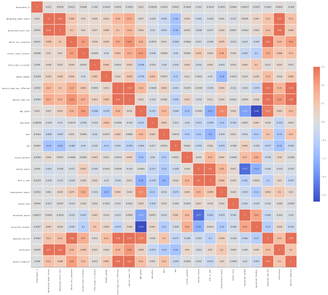

Figure 5: Heatmap after feature engineering

Standardisation

Standardisation to standard normal random variables with mean 0 and standard deviation 1:

$$z = \frac{x - \mu}{\sigma}$$

Motivation: features are in different scales, so standardisation brings them to comparable scales

Imperial College London

Outline

Imperial College London

Variable distributions

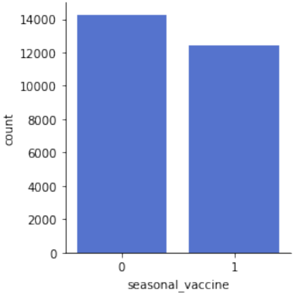

Figure 6: Outcome variable distribution

Approximately equal case-control distribution, meaning precision metrics (specificity and accuracy) won't be affected by balance

Variable distributions

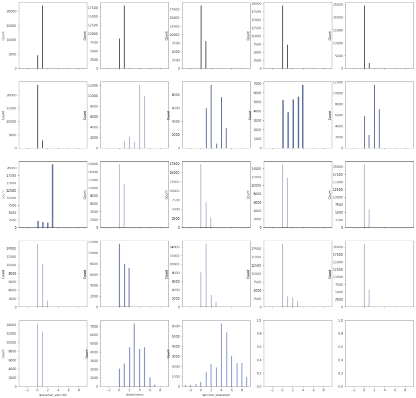

Figure 7: Predictor variable distributions

Outline

Imperial College London

Train and test splitting

80-20 train-test split

Motivation: 26'707 observations so a test set of 20% (small) will still give robust error metrics

Imperial College London

Outline

Imperial College London

Logistic regression

Motivation:

- Benchmark model
- Understand the predictive power of a simple model

Imperial College London

Logistic regression: performance

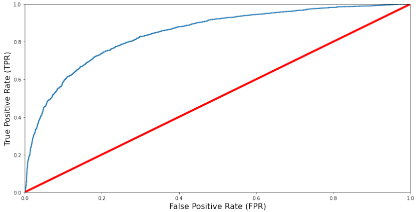

Figure 8: ROC curve
AUC= 0.84

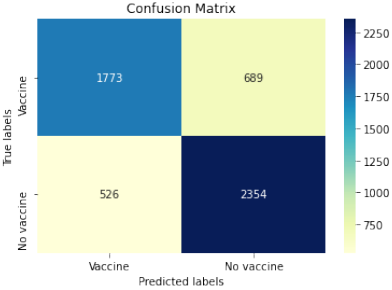

Figure 9: Confusion matrix
Sensitivity= 0.77
Specificity= 0.77
Precision= 0.72
Accuracy= 0.77
F1 score= 0.74

Outline

Random forest

Motivation:

- Understand the added accuracy of an ensemble algorithm
- Performs very well for classification problems compared to other improvements of decision tree algorithms
- Obtain a variable importance plot

Random forest: hyperparameter tuning

Hyperparameter	Base model	Tuned model
Estimators	100	1600
Min samples per leaf	2	4
Min samples per split	1	2

Table 1: Hyperparamters of base model vs new model tuned by cross-validated grid search

$\Rightarrow$ The new model was computationally intensive (80 min) but only improved AUC by 2$\Rightarrow$ Keep base model

Random forest: one tree

At the root node there are less samples than training data points highlighting the random forest's training with bagging (on a random subset with replacement)

Imperial College London

Random forest: variable importance plot

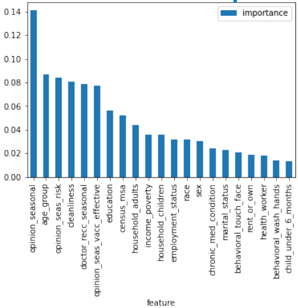

Figure 10: Variable importance plot

Relevance:

- Interpretability: view features most associated with outcome
- Reduce overfitting: remove non-contributing features

Imperial College London

Random forest: performance

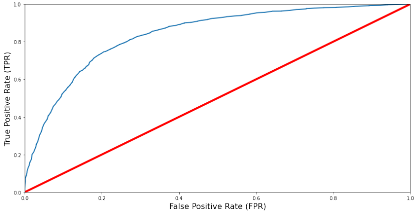

Figure 11: ROC curve
AUC= 0.84

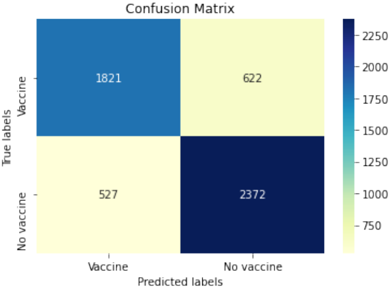

Figure 12: Confusion matrix

Sensitivity= 0.78

Specificity= 0.79

Precision= 0.75

Accuracy= 0.78

F1 score= 0.76

Imperial College London

Outline

Neural network: motivation

Motivation:

- Unsupervised learning algorithm
- Can handle complex data structure (21 features)

Imperial College London

Neural network: input and output layers

Input layer depth: 21 (number of unique features)

Output layer depth: 1 (binary classification with sigmoid activation function)

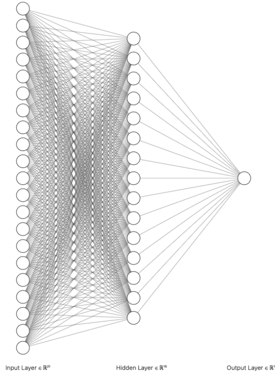

Figure 13: Neural network architecture

Neural network: hidden layer

Number of layers (depth):

- One hidden layer is sufficient for the majority of problems

Number of neurons in hidden layer (width):

- $<$ input layer width, $>$ output layer width
- $<$2 times input layer width
- $2/3$ input layer width $+$ output layer width

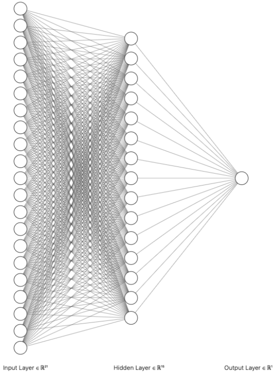

Figure 14: Neural network architecture

Imperial College London

Neural network: hyperparameter tuning

Method: Cross-validated grid search

Hyperparameters:

- Batch size: 100
- Epochs: 50
- Optimisation algorithm: Adam

Imperial College London

Neural network: performance

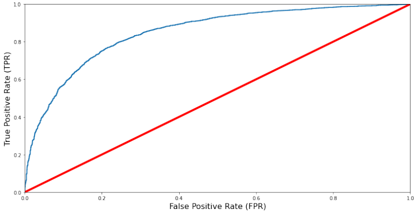

Figure 15: ROC curve
AUC= 0.85

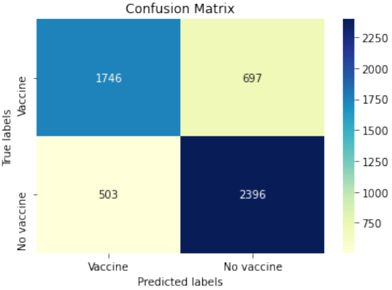

Figure 16: Confusion matrix
Sensitivity= 0.78
Specificity= 0.77
Precision= 0.71
Accuracy= 0.78
F1 score= 0.74

Imperial College London

Outline

Results

Model	ROC-AUC	Accuracy	F1 score
Logistic regression	0.845	0.77	0.74
Random forest	0.830	0.78	0.76
Neural network	0.85	0.78	0.74

Table 2: Performance metrics of three classification algorithms

$\Rightarrow$ All models have comparable performance metrics
$\Rightarrow$ Prediction performance by models is better than random chance, but is far from perfect
$\Rightarrow$ Measures are coherent, there is no over/under-representation of either class

Imperial College London

Outline

Discussion

$\Rightarrow$ Predicting vaccine uptake does not require complex transformations of data (linearly separable data)
$\Rightarrow$ However, perfect prediction is unlikely due to an emotive component to the nature of the decision which cannot be fully captured in the 21 features (in fact the biggest contributor to prediction is the opinion variable)

Discussion

Study relevance:

- Projecting vaccination rates
  - Informing decision of number of doses to purchase from manufacturers
  - Informing decision of how to allocated health resources
- Predicting whether immunisation programme will reach protective efficacy
- Leveraging factors most associated with vaccine compliancy

Imperial College London

Outline

Evaluation

Strengths:

- Research question could be answered with a simple model (logistic regression) that did not use too much computational power - a complex model is not always required
- Was able to choose a well-performing random forest over a very slightly better performing one to save compuatational power

Limitations:

- Could not tune all hyperparamters due to computational resources needed
- Data available cannot fully capture determinants of outcome due to the nature of the research question