**Machine Learning Project**

***Dataset:***

My original dataset contains over 26’000 observations for 36 features regarding three main areas: individual’s opinions on the flu vaccine (such as if it’s effective, if the flu is dangerous, if you can catch the virus from getting the vaccine), behaviours (if they wash their hands often, if they gather in large crowds, if they wear a face mask), and sociodemographic factors; and whether they took the seasonal flu vaccine or not as the outcome variable.

***Data pre-processing:***

***Missingness:***

I dealt with missing values by removing features which contained more than half of the observations missing, and for those features I kept I imputed the mode for categorical variables and the mean for numerical variables. There was also one column containing variables I could not decipher so I removed it.

***Categorical features:***

Because machine learning algorithms deal with numerical values, I encoded the categorical features. I used label encoding for hierarchical features and one-hot encoding for features that weren’t in a hierarchy, so that the algorithm wouldn’t assume that some levels are closer to each other than others.

***Collinear features:***

I saw in the heatmap which I will show in the next slide that there was a correlation structure in the features, so engineered two new features, one that encompassed all the behavioural features, and one that encompassed all the cleanliness features, that are listed on the slide, and checked that these two new features were approximately normally distributed as you can see in the histograms.

And here is the heatmap before and after feature engineering, showing that the block of high correlation in the upper left has been removed.

***Standardization:***

Lastly, I standardized all variables to standard normal variables in order to bring features to comparable scales.

***Exploratory analysis:***

***Variable distributions:***

I checked that the two classes of the outcome variable, whether individuals did or did not take the seasonal vaccine, were not too imbalanced. I did this to know which performance metrics to rely on the most after I fit my models, because an unbalanced distribution can skew especially specificity and accuracy.

I checked the distributions of all my features were approximately normally distributed in order for the algorithms to perform better.

***Train and test splitting:***

I split my dataset into 80% training and 20% testing. Why?

***Logistic regression:***

I started with a simple binary logistic regression as a benchmark model, to understand the predictive power of a simple model and later determine to what extent models that create more complex transformations of the data perform better.

This gave quite good discrimination with an AUC of 0.84, and very coherent although not perfect performance metrics.

***Random forest:***

I then used random forest to understand the added accuracy of an ensemble algorithm, and because it works particularly well for classification problems. I was also interested in a tree-based algorithm because I wanted to obtain a variable importance plot because I believe it is informative to the research question.

To run the random forest using SciKit learn on Python I first ran a base model with the parameters shown, and then I tuned those parameters using cross-validated grid-search. Tuning found that using 1’600 trees as opposed to just 100, and the other two new parameter values shown in the table, gave better performance. However, when I then computed the performance metrics for the tuned model, I found that it only improved the AUC by 2% (check this). Considering that it was much more computationally expensive than the base model since it took 80 minutes to run compared to less than one minute, I thought it was wise to stick with the base model.

As I said I was particularly interested in obtaining a variable importance plot, which shows the variables with the most decrease in node impurity. This is relevant to the research question because the features that are shown to be the most important in predicting whether individuals will take the flu vaccine are the ones that should be leveraged for an effective vaccination campaign. The most influent feature was the aggregate of all opinions on the seasonal vaccine (in the first column), and of those particularly the opinion on whether seasonal flu is dangerous (in the third column). Equally important was the aggregate of all cleanliness features, and whether doctor recommended the vaccine. Interestingly all sociodemographic factors apart from age group ranked low for variable importance.

However, when looking at the performance of the random forest classifier, it did not perform better than the logistic regression classifier, as every performance metric was nearly the same.

***Neural network:***

Lastly, I decided the fit a neural network to classify whether individuals will take the flu vaccine. I did this in order to experiment with an unsupervised learning algorithm, and because neural networks can deal with complex data structures by combining activation functions to best fit the data structure, and this could help with a dataset that has 21 features all regarding different aspects.

I decided to choose the size of the network manually using intuition. The input layer had a width of 21 because I had 21 input features, and the output layer had a width of 1 because I was performing classification with a sigmoid activation function. With regards to hidden layers, I chose a depth of one hidden layer because I found in the literature that one is usually sufficient, and I didn’t want to cause overfitting or an unnecessarily complex algorithm; I chose a width of that hidden layer of 15 by following the three rules listed below from the literature: Two-thirds the input layer width plus the output layer width; checking that this value lies in between input and output layer widths, and is less than twice the input layer width.

After having chosen the topology of the tree using intuition I tuned three hyperparameters using cross-validated grid search: the batch size, the epochs, and the optimization algorithm, which resulted in the values on the slide, and with those I fit my neural network.

Very surprisingly, also my tuned neural network did not improve performance compared to neither the logistic regression nor the random forest classifiers, improving AUC only by 1%.

***Results:***

Here is a summary table of the main three performance metrics for my three classification algorithms to predict whether individuals will take the seasonal flu vaccine. All models have comparable performance metrics, with a quite good discrimination between classes shown by the AUC of around 0.84, so much higher than a random classifier at 0.5. Additionally, all other performance metrics derived from the confusion matrix show very coherent results, meaning that there is no over- or under- representation of predictions for either class, so it is not more difficult for the model to predict a positive or a negative.

***Discussion:***

The fact that the prediction performance does not increase when using more complex models than a logistic regression, even with hyperparameter tuning, suggests that the prediction task given the data available does not require complex transformations of the data to replicate reality. Instead, the data might be quite well linearly separable.

Additionally, I believe that for the research question of predicting whether people will take the seasonal flu vaccine, there is a cap to how good the discrimination can get due to the emotive and somewhat irrational component in an individual’s decision. This is supported by the fact that the most important variable in the prediction according to random forest is the aggregate of all opinions, this means that yes we included some opinions that can be measured, but not all opinions are measurable or even explainable. In fact, the confusion matrix metrics show that there is no overrepresentation of neither predicted positives nor predicted negatives, simply suggesting a limit to the prediction performance that is possible. Considering this limit to how good the prediction can be due to the nature of the research question, I believe that an AUC of approximately 0.84 for the models makes a very good discriminator.

This study is relevant to public health in order to project annual vaccination rates against the seasonal flu. Projecting these rates is important to inform many decisions, such as deciding how many doses of the vaccine to purchase from the manufacturers, deciding how to allocate health resources if a particularly good or bad wave of influenza is predicted. Additionally, predicting how many people will take the vaccine will predict whether an immunization programme will reach the protective efficacy in the general population. Lastly, the factors most associated with the predictions can be leveraged to increase vaccine compliancy, these factors were found to be the opinion variables, suggesting that providing information about the seasonal flu is the most effective way to boost vaccination rates.

***Evaluation:***

I was satisfied in seeing that I was able to choose a well-performing random forest over avery slightly better performing one to save compuatational power, because that really familiarized me with the trade-off between computational expenditure and the extent to which hyperparameters can be tuned.