**Methodology**

To analyze each model, we used a similar structure to ensure each model was fitted appropriately. We followed these steps:

1. Split the original data into test and training sets.
2. Next we would check the model performance on the training set, using cross validation to tune the hyper parameters if necessary.
3. Find the accuracy on the test data.
4. Combine models using a weighted average (Voting Ensemble)

For the first step, we randomly sample and remove 20% of the original data to be used for testing. The test data is used to assess the model performance on data the model has not seen before. We assume the random sample has similar characteristics as the overall data.

For the second step, we choose the optimal parameters using the training data. Several models have other parameter that need to be optimized using the cross validation. When using cross validation, we split the data into several folds, usually 10 or 5, and fit the model and obtain the error rate with the suggested parameter to the data in the folds except one. We repeat this process and calculate the error rate for on each of the folds then find the average error rate. This process would result in the cross validation error rate for one parameter. We then repeat this cross validation process for the range of parameters we which to evaluate. Cross validation gives an accurate estimate of the model accuracy by reducing the in sample prediction variance.

In the third step, we evaluate the model with the optimal parameters from the training data on the test set. We hope the resulting accuracy will be similar to the prediction accuracy on data we do not have. We generally also compare this accuracy with the a priori (always guessing yes) estimate of the test set.

For the final step, we combined the models outputs to determine a final ranking. By combining models, the prediction ***variance*** would be decreased. We decided to give each model output a certain weight of the final output. We decided to have all the models have an equal vote except the random forest model which has 2. The random forest model is algorithmically different from the logistic regressions. Due to potential correlations between the logistic regression models, we wanted the random forest to have more slightly more power than the other models.