Potential Overfitting Problem

There are probably some problems with the stats I reported last time... I discovered this after visualizing the decision tree created by a decision tree classifier on the training set.

- The sampling approach I used (sample with replacement) can be dangerous if measures for overfitting are not taken, as points in the upsampled class gain more importance in the decision function.
- On top of this points can exist in both train and testing data -- especially dangerous for tree based classifiers, as if a path has been created for a point in training and the same point shows up in testing, the classifier is guaranteed to get it right

- No max depth with the random forest means that this could be a huge problem
- The Tree classifier (next slide), suggests that the above is a problem with the random forest classifier
- We might be looking at embellished stats for class 0 (not affected patients)
- However the tree might still be of interest to researchers

Performance and link to Decision Tree with no Pruning occurring

10-Fold CV Results:

Average Accuracy: 0.8

Stdev of Accuracy: 0.067

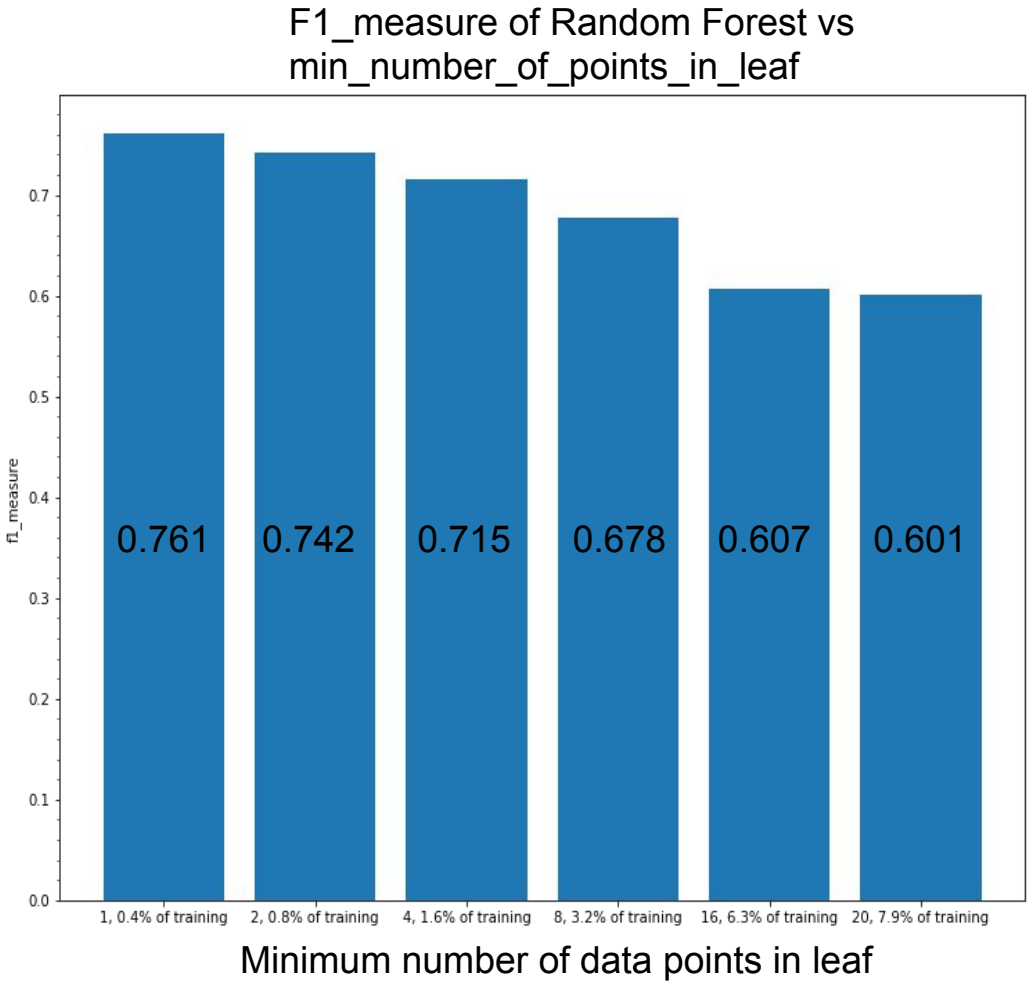
- Class 0 -> not affected
- Class 1 -> affected
- Class 0 was upsampled to fix the imbalance problem

For a single split, where test data is roughly 50% represented by each class, we have the following:

- AUC: 0.76
- Confusion Matrix,
  - [[**0.81** 0.19]
    [0.3 0.7]]
- Accuracy: 0.756
- Overall Precision: 0.759,
  - Class 0 Precision: 0.732
  - Class 1 Precision: 0.786
- Overall Recall: 0.755,
  - Class 0 Recall: **0.812**
  - Class 1 Recall: 0.698
- F1_Measure: 0.755

Link to Tree (from same split as below)

- If you open the tree diagram, you can see how precise the paths down to the leaves are
- While this isn't the forest, this tree is concerning enough to suggest things are not constrained in the forest classifier
    - The boosted performance of the forest classifier on class 0 further suggests that this is the case… class 0 is the upsampled class
- To the right you can see how the performance of the random forest degrades as you increase the minimum number of data points in each leaf (a measure to help with overfitting)



Addressing Gender Feature Importance Issue

- The Random Forest Classifier only gave 0.04 (out of 1) importance to gender as a feature
    - this trend continued with Tree Classifier, 0.06 importance
- This was surprising given the huge split between male and female protein levels for each reading.
- I think this might partly be because the other features are all numeric, while gender is binary. When deciding how to split the data, the tree is able to cycle through many more possibilities for the other features, while only 2 for gender.
- Also the splitting process is greedy, so at each individual level it might not be the best to split on gender.

To simulate what might happen if the tree(s) was forced to split on gender first, I will:

1. Split the original data (before the sampling) into male vs female
2. Run sampling on both sets to even out the data between affected and non_affected data points
3. Run the Random Forest and Tree classifier against the male and female data to see if there are any noticeable differences

Note: To keep parity I'm still allowing for overfitting

Random Forest Performance

Params: 30 trees used, no max depth, using gini coefficient to decide how to split trees

Female Data

Data = 63 affected, 30 not affected
Sampled-up the not affected class

Cross (10) Fold Mean Acc: **0.796**
Cross (10) Fold Stdev Acc: **0.082**

Averaging over 10 different train/test splits

mean_accuracy: **0.798**
median_accuracy: 0.786

mean_precision: **0.809**
median_precision: 0.825

mean_recall: **0.796**
median_recall: 0.799

mean_f1_measure: **0.793**
median_f1_measure: 0.782

Male Data

Data = 127 affected, 54 not affected
Sampled-up the not affected class

Cross (10) Fold Mean Acc: **0.764**
Cross (10) Fold Stdev Acc: **0.066**

Averaging over 10 different train/test splits

mean_accuracy: **0.761**
median_accuracy: 0.759

mean_precision: **0.776**
median_precision: 0.764

mean_recall: **0.765**
median_recall: 0.759

mean_f1_measure: **0.759**
median_f1_measure: 0.757

- No real difference between the performance of the classifiers on the two datasets, further suggesting that gender just isn't an important feature to the classifier… this isn't to say it's not actually important.
- Below you can see how the Random Forest performs with and without the gender feature.

Data: Data = 190 affected, 84 not affected
Sampled-up the not affected class

Without Gender

Cross (10) Fold Mean Acc: **0.796**
Cross (10) Fold Stdev Acc: **0.082**

Averaging over 10 different train/test splits

mean_accuracy: **0.798**
median_accuracy: 0.786

mean_precision: **0.809**
median_precision: 0.825

mean_recall: **0.796**
median_recall: 0.799

mean_f1_measure: **0.793**
median_f1_measure: 0.782

With Gender (Original Result)

Cross (10) Fold Mean Acc: **0.826**
Cross (10) Fold Stdev Acc: **0.078**

Averaging over 10 different train/test splits

mean_accuracy: **0.764**
median_accuracy: 0.768

mean_precision: **0.773**
median_precision: 0.782

mean_recall: **0.764**
median_recall: 0.771

mean_f1_measure: **0.761**
median_f1_measure: 0.766

- Again not much difference in classifier performance
- So in conclusion, it really does seem like the classifier doesn't think that gender matters too much when allowed to overfit to the data.

- When putting overfitting constraints on the learning process we have the following for feature importances:

Min_number_of_points	age_imp	**gender_imp**	BL_imp	V4_imp	V6_imp	V8_imp
1	0.18	**0.04**	0.25	0.2	0.16	0.18
5	0.15	**0.03**	0.25	0.21	0.16	0.19
9	0.13	**0.05**	0.26	0.23	0.18	0.15
13	0.1	**0.04**	0.33	0.18	0.16	0.19
17	0.09	**0.03**	0.25	0.25	0.17	0.22

- No real change in gender importance even when taking preventative overfitting measures.

Conclusion

I think that the greedy approach of the learning algo might be working against the gender feature in this case, along with the fact that there are so many more splits available for the other features.

In all I'm not sure how useful my original results were given the potential overfitting problem, but I thought that maybe they'd like to see the tree, as it might reveal some interesting thresholds for them.

Finally, if you still think that my findings are worth showing to researchers, here is the report you asked for… tree included.

Report