Report of Classification of RF and IJ Questions

Jerry Yin

July 10 2019

1 Classification of Risk Factors

To predict the risk factor based on this year's questions, I used bagging technique using three random forests as base classifier and use soft voting (voting based on weight):

- Random Forest (Gini Criterion)
- Random Forest (Entropy Criterion)
- Random Forest (without bootstrap)

Another boosting classifier is used to reinforce the result of bagging classifier, the number of iteration is set to be 10.

First, for the pairs of input, attitudinal questions, and output, risk factor of the next period. I did a census of the number of each label, and try to balance them in quantity.

Before Balance		After Balance	
Label	Number	Label	Number
0	3044	0	4044
1	1931	1	3862
2	1221	2	3663
3	996	3	2998
4	893	4	3572
5	845	5	3380
6	604	6	3020
7	361	7	2888
8	192	8	3264

Table 1: Balance of Labels

Then when the labels are balanced, the result is shown as following:

Classification Report of Bagging				
	Precision	Recall	F1	Support
0	0.76	0.86	0.80	701
1	0.83	0.85	0.84	759
2	0.95	0.94	0.94	722
3	0.96	0.93	0.95	595
4	0.98	0.96	0.97	712
5	0.99	0.95	0.97	728
6	1.00	0.97	0.98	638
7	1.00	0.98	0.99	621
8	1.00	1.00	1.00	661
Accuracy			0.94	6137
Macro Avg.	0.94	0.94	0.94	6137
Weighted Avg.	0.94	0.94	0.94	6137

The boosting has a better performance than the bagging classifier:

Classification Report of Boosting

	Precision	Recall	F1	Support
0	0.82	0.81	0.81	796
1	0.83	0.86	0.84	749
2	0.95	0.97	0.96	694
3	0.96	0.95	0.96	582
4	0.99	0.98	0.99	703
5	0.99	0.98	0.99	712
6	1.00	0.98	0.99	639
7	1.00	1.00	1.00	613
8	1.00	1.00	1.00	659
Accuracy			0.94	6137
Macro Avg.	0.95	0.95	0.95	6137
Weighted Avg.	0.94	0.94	0.94	6137

Generally, we can observe that the boosting classifier reinforced the machine's ability to distinguish people with higher risk factors, while its ability to distinguish people with lower risk factors is relatively low (like 80%). One possible reason is that originally we have less data with big labels. The part of those data probably has more distinguishable features, while data with small labels may be more similar and harder to distinguish.

In the testing set, we pick the first 50 data to show the result of the classifier. The real data is the following:

```
[8. 4. 2. 4. 2. 0. 5. 8. 8. 2. 3. 6. 6. 8. 5. 2. 1. 1. 1. 2. 2. 6. 5. 4.
 2. 1. 0. 0. 6. 1. 2. 5. 0. 8. 8. 2. 8. 6. 7. 0. 4. 7. 5. 3. 5. 8. 4. 1.
 3. 6.]
```

Figure 1: Real Label of the First 50 Instances in Test Set

The predicted data is the following:

```
[8. 4. 2. 4. 2. 0. 5. 8. 8. 2. 3. 6. 6. 8. 5. 2. 1. 1. 2. 2. 2. 6. 5. 4.
 2. 1. 0. 0. 6. 3. 2. 5. 0. 8. 8. 2. 8. 6. 7. 1. 4. 7. 5. 3. 5. 8. 4. 0.
 3. 6.]
```

Figure 2: Predicted Label of the First 50 Instances in Test Set

Comparing the labels, there are 4 mis-prediction, the accuracy is 92%. Also, all those mis-predictions are about labels from 0 to 3. Those children will be graduated from the program anyway, so the influence is not severe.

The whole training process takes 4.3743 seconds.

2 Classification of IJ Questions

The classification of ij Questions are harder than the classification of the risk factor. As the latter only has one output, while the former one has the output of 17 dimension (only considering the past 6 months questions).

The process of managing data and training is more complicated than that of the previous section:

- We need to balance the dataset by each question. Therefore, as the responses to ij questions are all binary questions, I calculated the ratio of 0 labels and 1 labels and adjusted them (separately for each dataset) to approximately the same;

- 34 bagging classifiers (hard and soft voting) based on Gaussian Naive Bayesian, Decision Tree, Random Forests are trained, and 17 boosting classifiers are trained based on the soft bagging classifiers;

- To test the accuracy, instead of using the "balanced dataset", I randomly picked 20% of the data from **original dataset**, and the test sample is considered accurate ONLY when ALL 17 sub-classifiers provide with the correct result.

The result of all sub-classifiers is too long to present, here is the average precision, recall, and accuracy for all the classifiers: (Note: time used is the time used by training all three classifiers)

Report of the Boosting Classifier				
No. of Classifier	Precision	Recall	Accuracy	Time Used (s)
IJ-40	0.98	0.98	0.98	4.52
IJ-41	0.98	0.97	0.98	5.46
IJ-42	1.00	1.00	1.00	4.84
IJ-43	0.92	0.88	0.89	3.84
IJ-44	1.00	0.99	0.99	5.69
IJ-45	0.98	0.98	0.98	5.95
IJ-46	0.99	0.99	0.99	4.73
IJ-47	1.00	1.00	1.00	4.95
IJ-48	0.99	0.99	0.99	4.84
IJ-49	1.00	1.00	1.00	6.15
IJ-50	0.99	0.99	0.99	4.99
IJ-51	1.00	1.00	1.00	5.25
IJ-52	0.96	0.96	0.96	4.69
IJ-53	1.00	1.00	1.00	4.95
IJ-54	1.00	1.00	1.00	5.03
IJ-55	1.00	1.00	1.00	5.16
IJ-56	1.00	1.00	1.00	5.02
Overall	0.988	0.984	0.985	86.09

In order to avoid over fitting by feeding to much self-copied data, I test the classifiers based on original data. I take 20% of the original data as testing set, and test the bagging and boosting classifiers on the set. The performance is the following:

Model	Accuracy
Hard Bagging	89.15%
Soft Bagging	77.45%
Boosting	92.12%

Table 2: Performance of the Three Classifiers

As I treat successfully predicting all 17 labels as accurate, the boosting classifier provides with a 92.12% accuracy, which is pretty high. If I relax the standard to successfully predicting 16 of 17 labels, then the accuracy of the boosting classifier rises to 98.61%, and the other two have the accuracy of 99.11% and 94.10%.