Abalones dataset – unbalanced data set

10/6/2016 10:11 PM – **Try1.R** has a random forest thrown at the dataset. I knew it would not work but I wanted to establish a baseline for my other algorithms and tests.

In the notes they mention they use a training set of 3133 and a test set of 1044.

I wasn’t sure if it was just the first 3133 observations or if it should be randomly sampled. When just the first 3133, the model overfits and I think I get a perfect prediction. (This must not be the case)

In the second half of the code, I use a random sample and realize that my baseline accuracies for the **training set is 25.7%** and **my test set is 22.5%**

**New Idea**

1. Split the data set into datasets with similar class sizes.
2. Fix the each local dataset class imbalance problem
3. Run algorithm on each dataset and attempt to get a probability of being in a class
   1. Run multiple algorithms- ensemble to get a probability- when ensembling we can tweak each local dataset’s algorithms weights
4. Compare probabilities
   1. WE may need to do final adjustment for data set size and variance for the probabilities
5. Choose algorithm with the largest probability
6. How to split the datasets up-

6 classes 267 – 689 observations

|  |  |
| --- | --- |
| Class | Number of Examples |
| 7 | 391 |
| 8 | 568 |
| 9 | 689 |
| 10 | 634 |
| 11 | 487 |
| 12 | 267 |

4 classes 103 – 203 observations

|  |  |
| --- | --- |
| Class | Number of Examples |
| 5 | 115 |
| 13 | 203 |
| 14 | 126 |
| 15 | 103 |
| 16 | X< |

6 classes 26 – 67 observations

|  |  |
| --- | --- |
| Class | Example |
| 4 | 57 |
| 16 | 67 |
| 17 | 58 |
| 18 | 42 |
| 19 | 32 |
| 20 | 26 |

11 classes, 1 - 15 observations

|  |  |
| --- | --- |
| Class | Example |
| 1 | 1 |
| 2 | 1 |
| 3 | 15 |
| 21 | 14 |
| 22 | 6 |
| 23 | 9 |
| 24 | 2 |
| 25 | 1 |
| 26 | 1 |
| 27 | 2 |
| 29 | 1 |

10/6/2016 10:11 PM- Started working on SMOTE for first data set. It is a little bit strange because it changes the size of the majority classes as well as the minority class. I am not quite sure how it works but I found a parameter that makes the classes similar to the previous size. Hopefully it will keep the original observations

I will Smote 2x so that the class sizes will be even closer. They will all be between 15-20% per class.

I am concerned about bias and distributional changes to the variables. I want to compare the new datasets to the original. I want to check the mean, variance and quantiles (5,25,50,75,95) for each variable and see the percentage change. Just so I know the algorithm isn’t adding too much bias to the variables…

But that’s a project for the next day.

I wish there was some algorithm for how to pick the sizes. Like 100 % and then how to find the perc.under parameter value. Thoughts for another day (# of observations + minority sampling class) .

On to using algorithms

Ran a random forest , class error is less than ~10% for each class