Assign3 writeup

Santina

Sunday, November 05, 2014

In this assignment, we examined a few different linear models that are trained based on different attributes or on data that were imputed.

The source code is loaded here.

```
source("Assign3_SantinaLin.R")
```

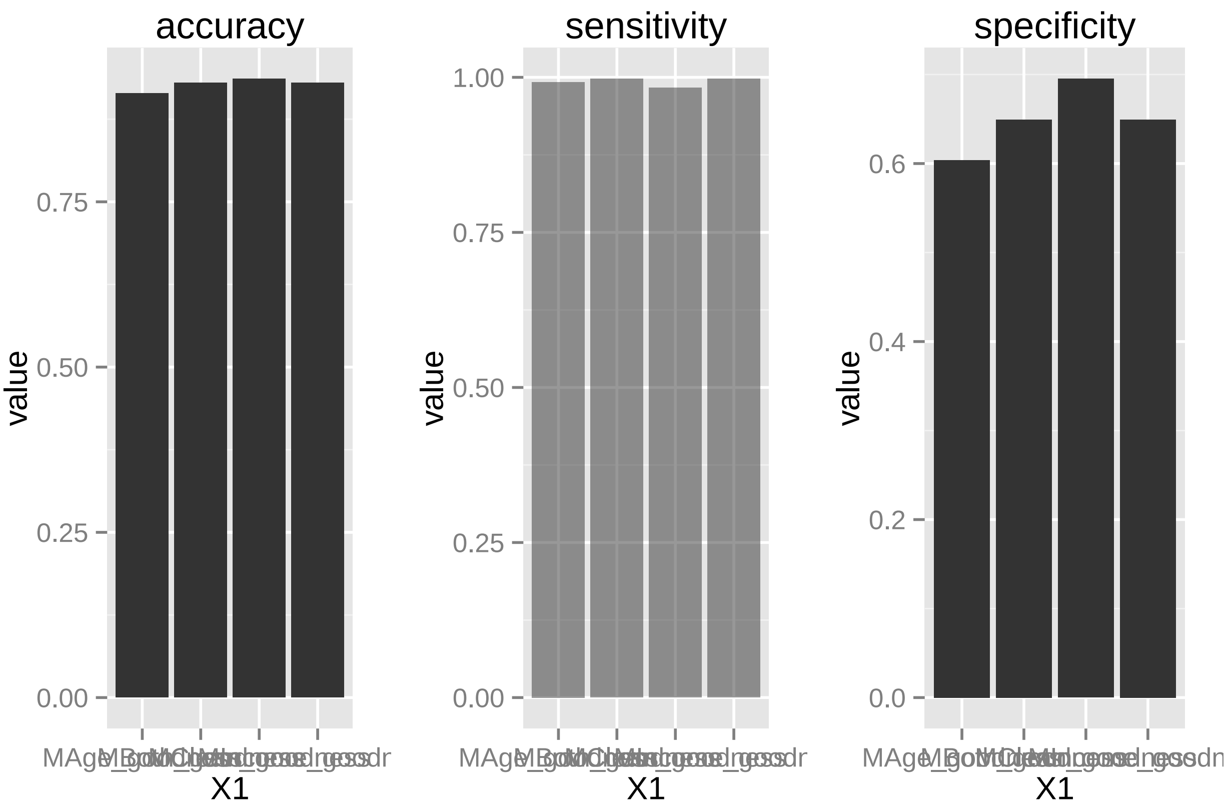
And here's the summary of all four models in comparison.

```
summary
```

```
##                    accuracy sensitivity specificity
## MClean_goodness       0.936      0.9833      0.6951
## MBoth_goodness        0.930      0.9975      0.6495
## MIncome_goodness      0.930      0.9975      0.6495
## MAge_goodness         0.914      0.9925      0.6040
```

Let's look at them in graph

```
grid.arrange(accuracy_summary, sensitivity_summary, specificity_summary, ncol=3)
```



We can see that MClean has the highest specificity, accuracy, but a slightly lower sensitivity. MClean was trained on the cleaned data set with no missing value, whereas the other three models are trained on the cleaned data set plus the imputed missing values. We can see that using this method to increase the number of training samples don't always help with the accuracy of a trained model. With the exception of sensitivity, the trained model on imputed data values perform slightly worse than MClean.

The smallest error does correlate to how well a model performs, based on the accuracy. The predictions made based on age has the highest error out of all three models, and it has the lowest accuracy, sensitivity, and specificity out of all three models trained on imputed values. Models trained on income or both age and income give better results. Therefore, how closely predicted imputed missing values to the true values have an effect on how well a model, trained on the imputed values, can perform, assuming the true values correlate with the prediction on loans.