|  |  |
| --- | --- |
| **Criteria** | **Points (0-10)** |
| **Content** | 10 |
| **Subject Knowledge** | 7 |
|  | 17/20 |

**General Feedback**

**Content**

* Repo excellently organized with relative path names in all scripts
* Very neat code with comments throughout
* Efficient code, like the use of separate scripts per analysis section
* All tables generated by R as requested, complete citations given (nice package citations!)

**Subject Knowledge**

* Great background on importance of stroke prediction, though would like more info on the cited meta-analysis of past prediction works (what was the accuracy of the SVM and RF algos?)
* Good detail on sample size and missing data
* Good mention of SMOTE and use due to imbalanced data (what were your over/undersampling settings, it is in your script but should be mentioned here as well for reproducibility. Also should mention what variables you used for SMOTE, i.e. what was on the other side of the ~. Should also mention that you correctly used SMOTE only on the training set in your analysis, with the test fold untouched.
* Good selection of algos, though why no ROC analysis? Would add to your analysis pipeline to compare different probability thresholds. Also, you used a 0.5 threshold for RF as well (which is ok, just making sure you understand this as it is not explicit in R) which means you could have done a ROC analysis with RF as well.
* Good detail on tuning, though maybe mention what metric was used for tuning to select the best parameters (I’m assuming accuracy, but should be explicit). Also based on your code, it looks like out-of-bag error was used to tune for the RF and not 5-fold CV (which is perfectly fine!).
* Good summary of summary stats results. You could tie these differences in the sample in Table 1 to the variables that were most useful for prediction in the different algorithms, specifically looking the betas from the logistic regression and variable importance index/Gini change from RF. Would strengthen the results by adding some context to the classifiers. Also should have sample sizes for the stroke/no stroke groups included.
* Table 2 could be formatted better, specifically replacing “Class 0” and “Class 1” with the group labels related to stroke, as well as removing the “\_” and capitalization being more standard.
* Would like to see discussion on how your results compared to the ones in the cited paper, specifically for SVM and RF
* Analysis as a whole is good, though a bit basic. Could use a visualization to make the results pop more, and some additional analyses to add more context to the results.