Week_11_Assignment
Week 11 - Comparison of R/Python for XGBoost
HDS-5230-07
Shreya Samir Kurani

1. **Compare the accuracy values of XGBoost models fit on the newly created data, for the following sizes of datasets. Along with accuracy, report the time taken for computing the results. Report your results in a table with the following schema.**

METHOD USED	DATASET SIZE	TESTING-SET PREDICTIVE PERFORMANCE	TIME TAKEN FOR THE MODEL TO BE FIT (SECONDS)
XGBOOST IN PYTHON VIA SCIKIT-LEARN AND 5-FOLD CV	100	0.94	0.33
	1000	0.954	0.42
	10000	0.958	0.84
	100000	0.962	7.31
	1000000	0.964	75.68
	10000000	0.963	706.32
XGBOOST IN R – DIRECT USE OF XGBOOST() WITH SIMPLE CROSS-VALIDATION	100	0.98	0.02
	1000	0.998	0.04
	10000	0.9856	0.06
	100000	0.9707	0.4
	1000000	0.9672	1.81
	10000000	0.9663	19.9
XGBOOST IN R – VIA CARET, WITH 5-FOLD CV SIMPLE CROSS-VALIDATION	100	0.9505	1.53
	1000	0.9460	2.64
	10000	0.9555	1.89
	100000	0.9556	9.22
	1000000	0.9540	86.14
	10000000	Training not feasible due to hardware constraints	N.A

2. Based on the results, which approach to leveraging XGBoost would you recommend? Explain the rationale for your recommendation.

From the comparison of the three approaches, I suggest to train the model using XGBoost in Python using scikit-learn and 5 fold CV. All dataset sizes were consistently predicted with high performance keeping excellent trade off between accuracy and computational cost with this strategy. Computation times were slower for Python integration with scikit-learn than when performing the explicit application of xgboost() in R and training with caret, but at the expense of little predictive accuracy gain. Additionally, Python's platform provides richer degrees of flexibility and scalability for further optimization and deployment, and thus this is more practical for machine learning pipelines on very large scales.

Apology and Clarification

I initiated the training of the XGBoost model on the dataset that consists of 10 million rows using the caret package and it ran for over 20 minutes. The training ran for a very long time, but it never finished or produced any output, which led to a very high risk of system crash or instability. Hence, the training was manually stopped in order to maintain system integrity and avoid potential data loss. Thus, results for the 10 million dataset size are omitted specifically for the caret based approach. The model performance trends were found to be consistent for datasets of up to 1 million rows, and complete evaluations were successfully conducted. Additionally, since GitHub's file upload size is 25MB, synthetic dataset file synthetic_data_10000000.rds could not be uploaded to repository. However, the dataset was created from the dataset successfully and was used for local analysis and the outcomes of which have been integrated into the report.