| **Method used** | **Dataset size** | **Testing-set predictive performance** | **Time taken for the model to be fit** |
| --- | --- | --- | --- |
| XGBoost in Python via scikit-learn and 5-fold CV | 100 | 0.9 | 0.13 |
| 1000 | 0.965 | 0.06 |
| 10000 | 0.9745 | 0.15 |
| 100000 | 0.9872 | 0.83 |
| 1000000 | 0.9918 | 9.94 |
| 10000000 | 0.9932 | 103.23 |
| XGBoost in R – direct use of xgboost() with simple cross-validation | 100 | 0.8966 | 1.08 |
| 1000 | 0.9365 | 2.14 |
| 10000 | 0.9723 | 4.59 |
| 100000 | 0.9826 | 17.43 |
| 1000000 | 0.9859 | 128.53 |
| 10000000 | 0.9873 | 1424.53 |
| XGBoost in R – via caret, with 5-fold CV simple cross-validation | 100 | 0.8966 | 1.76 |
| 1000 | 0.9298 | 2.59 |
| 10000 | 0.9763 | 4.39 |
| 100000 | 0.9830 | 37.14 |
| 1000000 | 0.9852 | 130.33 |
| 10000000 | 0.9869 | 1075.81 |

The XGBoost implementation through scikit-learn achieves the best predictive results across all dataset sizes by reaching accuracy scores of up to 0.9932 when processing 10M observations. The implementation demonstrates excellent computational speed because it processes the largest dataset (10M) within 103.23 seconds while the R implementations needed 1424.53 seconds and 1075.81 seconds to complete the same task. The Python implementation demonstrates consistent superior performance across all data scales because it scales efficiently with minimal time increases when data volumes grow tremendously.

The predictive outcomes of both R implementations appear satisfactory yet their computing performance becomes highly inefficient as the dataset increases in size. The R implementation of xgboost() produces slightly better execution times for smaller datasets yet requires 14 times longer processing for the 10M dataset compared to Python. The caret approach in R provides better prediction results than xgboost() direct implementation yet delivers results behind Python execution and needs tenfold longer execution duration for large datasets. Longer time requirements in production settings where model retraining or inference needs to be performed on extensive datasets would become a major challenge because of this time difference.