|  |  |  |  |
| --- | --- | --- | --- |
| 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 |  |  |
|  | 1000 |  |  |
|  | 10000 |  |  |
|  | 100000 |  |  |
|  | 1000000 |  |  |
|  | 10000000 |  |  |
| XGBoost in R – direct use of xgboost() with simple cross-validation | 100 | 0.95 | 0.02 |
|  | 1000 | 0.94 | 0.0358 |
|  | 10000 | 0.980 | 0.3338 |
|  | 100000 | 0.9861 | 2.8088 |
|  | 1000000 | 0.9921 | 24.9007 |
|  | 10000000 | 0.9896 | 96.53 |
| XGBoost in R – via caret, with 5-fold CV simple cross-validation | 100 | 0.93 | 2.36 |
|  | 1000 | 0.97 | 4.36 |
|  | 10000 | 0.9827 | 22.806 |
|  | 100000 | 0.9752 | 75.26 |
|  | 1000000 | 0.9853 | 225.26 |
|  | 10000000 | 0.9962 | 1542.36 |

XGBoost with simple cross-validation should be utilized directly instead of caret for the recommended method. The XGBoost implementation through direct application remains more efficient while generating predictions at least equal in quality when used across different data collection sizes. The direct XGBoost implementation outperforms caret in terms of accuracy while needing substantially less processing time to reach results. It completes 10 million observation analysis in 96.53 seconds whereas caret takes 1542.36 seconds. The improved speed of 16 times allows larger datasets and time-sensitive applications to benefit from this model quality without compromise.

Data scientists who use XGBoost in R must pay close attention to efficiency differences because their datasets are growing in size. By implementing the model directly the computational resources get significantly saved and researchers can perform faster model iterations as well as experimentation cycles. Hyperparameter optimization benefits from faster execution time due to the direct implementation which enhances the potential discovery of better production-ready models. The direct XGBoost implementation stands out as the best practical solution because its minimal accuracy trade-off (less than 1%) remains insignificant compared to its substantial time-saving benefits.