Comparing Predictive Analytics Libraries based on Historical Datasets

András Béleczki[a], **Zoltán Vincellér**[b]

[a]Eötvös Loránd University
beleczki.andras@outlook.com

[b]Eötvös Loránd University
vzoli@inf.elte.hu

Abstract

Nowadays every company on the market realizes the importance of data science, since its primary goal is to help them make more informed decisions. This type of projects and tasks need another kind of developers: data scientists, predictive modelers and other mathematical-analytical professionals. There are a lot of tools and libraries that are essentials for specialists or at least can ease their work. Because the Big Data task are still very resource intensive most of the above mentioned tools are implemented in a cloud-environment.

In the case of cloud services there is a tense competition, where they can be only competitive if they turn the developers on their sides. This can be acomplished by providing easy-to-use methods, accurate algorithms and high performance computing. One of these solutions is the SAP HANA database[1], which is a column-based in-memory database system. It provides a lot of tools besides the Predictive Analytics Library, and its newest release (v2) came out with many performance improvement and features. Our purpose is comparing this architecture's capabilities with its previous version and with another popular solution: a Python library called Pandas[2]. This comparison is based on large volume of historical datasets provided by various sources.

Keywords: Big Data, Predictive Analytics, SAP HANA, Python Pandas

References

[1] Färber, F., Cha, S. K., Primsch, J., Bornhövd, C., Sigg, S., Lehner, W., SAP HANA database: data management for modern business applications, *ACM Sigmod Record*, (2012), 45–51.

[2] McKinney, W., pandas: a Foundational Python library for data analysis and statistics. , *Python for High Performance and Scientific Computing*, (2011), 1–9.