Publication date:
17 Feb 2021
Author:
Bradley Shimmin
Chief Analyst, AI & Data Analytics

Oracle democratizes machine learning with Database 21c release

Brought to you by Informa Tech

Table of Contents :

Omdia view

Summary

Oracle has released a new iteration of its flagship database management system, Oracle Database 21c, which incorporates among many useful data-centric enhancements, an interesting twist, a data science tool designed to automate the creation of machine learning (ML) models by (just about) anyone in the enterprise.

The role of the database in an ML workflow

Typically, specialization and refinement drive forward momentum for products as they evolve over time. Most technology companies seek out new market opportunities through expansion or acquisition, building or buying into new use cases with new products and services. But Oracle is not a typical technology provider. While the company certainly and aggressively builds and buys new offerings like any other vendor, it must also innovate within the confines of one single product, a product which is perhaps its most important asset as it underpins much of its portfolio -- Oracle Database.

In a way Oracle operates as a steward, a caretaker for Oracle Database. It must ensure that this single product (Oracle refers to it as a converged database), which is in its fourth decade now, evolves to meet the ever-changing market landscape for data storage and processing. The trouble is that key trends driving the enterprise database market, most notably artificial intelligence (AI) and ML, operate at a distance from the database. Data scientists building AI outcomes in the enterprise typically work within highly iterative environments using tools such as Jupyter notebooks and development languages such as Python and R, drawing not from a single datastore but rather from a vast array of data sources that evolve according to demands.

To data scientists, the database engine represents an external resource that can act as more of an impediment than an asset in terms of accessibility, quality, and performance. Enterprise AI practitioners often find themselves jumping through hoops to gain access to remote data sources, fighting against data model/format disparities, and struggling to access data in a continuous and performant manner. These and many similar operational challenges have fueled the rapid rise of ML operationalization (MLOps) platforms with specialized solutions such as Amazon SageMaker Suite and Dataiku Data Science Studio (DSS) introducing ML workflow capabilities that streamline access to data across the ML lifecycle. However, even these solutions still presume an arms-length relationship between data science and data management.

This leads us back to Oracle's stewardship of Oracle Database. Instead of chasing new database market opportunities through bifurcation (eg. adding more databases to meet specific data requirements), Oracle has focused its energies on evolving Oracle Database into a flexible, comprehensive data layer, spanning both analytical and transactional workloads while supporting a disparate collection of data models and data types -- an approach Oracle defines as a converged database architecture.

As an example, in August 2020, Oracle introduced Oracle Autonomous JSON Database, a specialized JSON document store. Running on Oracle's flagship cloud database, Oracle Autonomous Database, this new offering made for an affordable and easy-to-use entry point for developers specializing in JavaScript Object Notation (JSON) development. Oracle's intent is to create a single data fabric that can simultaneously and equally support spatial data, document stores (like JSON), streaming data (for IoT for devices), real-time in-memory analytics processing, as well as traditional relational data.

Introducing Oracle AutoML

It should come as no surprise then to see Oracle begin building advanced analytics capabilities into Oracle Database in support of data science workloads. For almost two decades the company has been working to build intelligence into its flagship database. In January 2021, Oracle took another sizable step in this direction with the introduction of Oracle Database 21c. Running on Oracle Cloud (and available on the company's free tier of Oracle Autonomous Database), this new release features a number of data-centric enhancements including blockchain tables, in-database JavaScript execution, in-memory graph analysis capabilities, and native binary JSON document support.

It also introduced what at first blush may appear as a strange addition to those not familiar with the evolution of Oracle's database, the inclusion of automated ML (AutoML). In fact, however, this addition directly builds on Oracle's existing efforts to support in-database ML algorithms, work that have been ongoing since the introduction of Oracle Database since 9i in 2001. With this release, for example, Oracle incorporated data mining and advanced analytics capabilities, featuring more than 50 statistical functions (regression, classification, even decision trees).

Popularized by Google in 2018 as a means of speeding up the task of selecting the most appropriate neural network to complete a given deep learning (DL) task, AutoML has grown into a more widely applicable means of automating a wide array of ML tasks including data preparation, model selection, feature selection and engineering, as well as hyperparameter tuning. Touted as a means of democratizing data science, AutoML allows non-technical domain experts to participate in the data science process. And true to its roots, it still allows data scientists to speed up and standardize repetitive tasks. These tools can be found within broader MLOps platforms as with Amazon SageMaker AutoPilot, as a stand-alone tool like H2O.ai AutoML, or as a pure Python framework like Auto Sklearn.

With Oracle Database 21c, Oracle enters this space with an AutoML solution that seeks to support both data scientists and non-technical domain experts. Available as a feature of OML4Py, which is itself a component included with Oracle Database and Oracle Autonomous Database, the company's new AutoML feature focuses on three main tasks: algorithm selection, feature selection, and hyperparameter tuning. And like most AutoML tools, it endeavors to both improve data scientist productivity and enable non-experts to participate in data science.

Embedding ML within the database itself

Interestingly, OML4Py and OML AutoML UI do not live within an overarching MLOps platform as with most AutoML solutions, but rather (consistent with Oracle's converged strategy) within Oracle Database itself. Why should enterprise AI practitioners run ML workflows inside of a database? As with real estate, the answer is location, location, location. Compared with orthodox ML tools that sit alongside data sources, in-database ML tools like OML4Py that leverage Oracle Machine Learning's in-database algorithms minimize the movement of data, allowing ML algorithms to transform and process data in situ. This confers a number of principal benefits compared with traditional, external Python data interfaces:

- Streamlines data access as practitioners can work directly with data without having to import data as a flat file or request data extracts from IT.
- Cuts down on data latency associated even direct API calls from database to Python engine.

- Bypasses memory limitation issues where the Python processing engine cannot load all requested data into memory before processing that data.
- Avoids the performance issues associated with single-threaded Python processes and ML algorithms by allowing them to run in parallel.
- Limits security vulnerability by removing external database authorization requests. This also eliminates the need to extract data to flat files that then get downloaded to local machines.

OML4py with AutoML is not alone in putting these in-database benefits to work. As a matter of fact, Oracle has been quite busy recently, creating a library of ML-specific database tools, Oracle Machine Learning (OML) covering a broad set of data science tasks.

- **OML4SQL** – an API and user experience that allows practitioners to use over 30 ML algorithms and score data using in-database models directly in the database using SQL.
- **OML Notebooks** – a notebook user experience built on Apache Zeppelin, which can run not just Python but also SQL and PL/SQL. Note that R language support is coming soon.
- **OML4R –** an R API to scalable native Spark-based ML algorithms and integrated MLlib algorithms, and a user experience (via Oracle Data Miner) that allows practitioners to access and manipulate data directly in-database via the R programming language and deploy user-defined R functions in database spawned and controlled R engines, invocable from SQL.
- **Oracle Data Miner –** an extension to Oracle SQL Developer, which enables data analysts and data scientists to create, test, and share ML methodologies from a drag-and-drop user interface.
- **OML4Spark –** an R API to scalable native Spark-based ML algorithms and integrated MLlib algorithms supporting big data sources such as functions HDFS, HIVE, Spark DataFrames, etc.
- **OML4Py –** a Python API that allows practitioners to access and manipulate data directly in-database via the Python programming language; this API also allows developers to deploy user-defined Python functions in database spawned and controlled Python engines, invocable from REST endpoints. Like OML4R, it enables users to leverage Oracle Database as a high performance computing (HPC) platform for data science.
- **OML AutoML UI –** built on top of OML4Py this user interface (UI) enables no-code ML model creation directly within Oracle Autonomous Database while supporting model deployment to OML Services.
- **OML Services –** a REST API-based model deployment and management service, with cognitive text analytics.

Note that AutoML, OML4Py, and OML Services are all now available to customers. Also note that when AutoML UI is released it will only be available only within Oracle Autonomous Database. To Oracle's credit the company has chosen to position its suite of ML tools not as separate products with an up-charge for each. Rather, Oracle Database owners and subscribers can make use of these tools within the current scope of their Oracle contracts. Further, Oracle is making all of these tools available within its free tier on Oracle Cloud (with, including AutoML. In this way, Omdia believes Oracle will encourage enterprise customers,

particularly those seeking to bring non-technical domain experts into the data science process, to look at Oracle Database as something more than a simple data "source" for data science.

Scope of operation for OML AutoML UI

In terms of functional capability, Oracle's new AutoML solution is very much in line with most AutoML tools, automating the selection of ML algorithms, the processing and selection of the features that will be used by the selected algorithm, and the refinement of the parameters (hyperparameters) used in tuning the resulting data model. These three steps can be operated code-free using Oracle's OML AutoML UI with users selecting the desired data source (e.g. a table or view in Oracle Database), picking the target variable and optionally the type of prediction (classification or regression), and then fine-tuning the experiment, selecting the number of models, training duration, accuracy measure, and specific algorithms to apply. Note that AutoML UI will automatically choose classification or regression depending on the datatype of the target. Regardless, users can choose to override any of these automated configuration options.

Once set in motion, OML AutoML UI tool will display a simple leaderboard, highlighting the progress of active evaluations along with the accuracy scores for each selected model type (random forest, generalized linear model, neural network, etc.) according to the selected measure (F1, recall, R2, etc.). Once the current run has completed, the AutoML UI will show users the ranking of the features that were of the most importance in predicting the target.

Once users are happy with the results of a given run (called an experiment), they can then deploy a completed model as an OML Services endpoint. This endpoint can then be shared with other users for further experimentation or made available in production via a RESTful API request. Note that many of the capabilities within OML AutoML UI are available within OML4Py for users who prefer to work from within a Python-based notebook environment. This is important because users opting to start with the code-free UI can transition their work to a notebook environment for further development and deployment. OML AutoML UI also generates notebooks with the hyperparameters used to build specific models, which also enables further model refinement.

Future avenues of exploration for Oracle

Compared with many pure play AutoML tools such as H2O.ai AutoML and open source Python libraries like Auto-Sklearn, Oracle's implementation has plenty of room to grow. There's an opportunity to more fully automate and augment data preparation and exploratory data analysis (EDA) functions such as imputation, one-hot encoding, binning, and standardization. There's also an opening for Oracle to extend the set of algorithms included in AutoML to support techniques like time series forecasting, growing beyond tabular data to also support unstructured data, for example. Given the many performance, access, and security benefits of doing ML in-database mentioned above, this multi-modal notion in particular could help Oracle differentiate within an exceptionally crowded market of AI platform providers.

Further the company can expand its algorithm portfolio to include advanced ensemble techniques (where multiple algorithms are combined) beyond current support for partitioned models, which is a type of ensemble. And while Oracle's AutoML feature provides a degree of explainability and transparency, documenting important metadata elements such as feature scores, the company should explore further

means of engendering trust among non-data scientists by providing users with a "what if" capability to actively visualize how experiment modifications might impact experiment outcomes.

Even with these potential roadmap items outstanding, Oracle has, with this first iteration of AutoML, definitely hit on all of the major tasks involved in the creation and deployment of regression and classification ML models. More importantly, Oracle can use this new feature along with its broader suite of in-database OML portfolio to make an important point with both existing and potential customers. Oracle Database 21c (and by extension Oracle Autonomous Database) is not an arms-length data resource for data scientists as is the case with specialized ML tools. Oracle users can uniquely run ML workloads internally in Oracle Database, accommodate a broad range of programming and query languages, support disparate data models and workloads, and with the introduction of AutoML create opportunities for enterprise practitioners, both data scientists and business domain experts.

Appendix

Further reading

"Analyst Commentary: Splice Machine aims to meld relational databases with machine learning," (April 2020)

"Oracle spins up an autonomous JSON document database," (August, 2020)

"Oracle looks to take the "Oops!" out of data science," (February, 2020)

Author

Bradley Shimmin, Chief Analyst, AI Platforms, Analytics, and Data Management

askananalyst@omdia.com

Citation policy

Request external citation and usage of Omdia research and data via citations@omdia.com.

Omdia consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help you. For more information about Omdia's consulting capabilities, please contact us directly at consulting@omdia.com.

Copyright notice and disclaimer

CONTACT US

omdia.com

askananalyst@omdia.com