Score! Team NVIDIA Takes Trophy in Recommendation Systems

A five-person global team recounts how it won a prestigious challenge in recommenders, the AI engines of the digital economy.

Author: Rick Merritt

A crack NVIDIA team of five machine learning experts spread across four continents won all three tasks in a hotly contested, prestigious competition to build state-of-the-art recommendation systems .

The results reflect the group’s savvy applying the NVIDIA AI platform to real-world challenges for these engines of the digital economy. Recommenders serve up trillions of search results, ads, products, music and news stories to billions of people daily.

More than 450 teams of data scientists competed in the Amazon KDD Cup ‘23 . The three-month challenge had its share of twists and turns and a nail-biter of a finish.

Shifting Into High Gear

In the first 10 weeks of the competition, the team had a comfortable lead. But in the final phase, organizers switched to new test datasets and other teams surged ahead.

The NVIDIANs shifted into high gear, working nights and weekends to catch up. They left a trail of round-the-clock Slack messages from team members living in cities from Berlin to Tokyo.

“We were working nonstop, it was pretty exciting,” said Chris Deotte, a team member in San Diego.

A Product by Any Other Name

The last of the three tasks was the hardest.

Participants had to predict which products users would buy based on data from their browsing sessions. But the training data didn’t include brand names of many possible choices.

“I knew from the beginning, this would be a very, very difficult test,” said Gilberto “Giba” Titericz.

KGMON to the Rescue

Based in Curitaba, Brazil, Titericz was one of four team members ranked as grandmasters in Kaggle competitions, the online Olympics of data science. They’re part of a team of machine learning ninjas who’ve won dozens of competitions. NVIDIA founder and CEO Jensen Huang calls them KGMON (Kaggle Grandmasters of NVIDIA), a playful takeoff on Pokémon.

In dozens of experiments, Titericz used large language models (LLMs) to build generative AIs to predict product names, but none worked.

In a creative flash, the team discovered a work-around. Predictions using their new hybrid ranking/classifier model were spot on.

Down to the Wire

In the last hours of the competition, the team raced to package all their models together for a few final submissions. They’d been running overnight experiments across as many as 40 computers.

Kazuki Onodera, a KGMON in Tokyo, was feeling jittery. “I really didn’t know if our actual scores would match what we were estimating,” he said.

Deotte, also a KGMON, remembered it as “something like 100 different models all working together to produce a single output … we submitted it to the leaderboard, and POW!”

The team inched ahead of its closest rival in the AI equivalent of a photo finish.

The Power of Transfer Learning

In another task, the team had to take lessons learned from large datasets in English, German and Japanese and apply them to meager datasets a tenth the size in French, Italian and Spanish. It’s the kind of real-world challenge many companies face as they expand their digital presence around the globe.

Jean-Francois Puget, a three-time Kaggle grandmaster based outside Paris, knew an effective approach to transfer learning . He used a pretrained multilingual model to encode product names, then fine-tuned the encodings.

“Using transfer learning improved the leaderboard scores enormously,” he said.

Blending Savvy and Smart Software

The KGMON efforts show the field known as recsys is sometimes more art than science, a practice that combines intuition and iteration.

It’s expertise that’s encoded into software products like NVIDIA Merlin , a framework to help users quickly build their own recommendation systems.

Benedikt Schifferer, a Berlin-based teammate who helps design Merlin, used the software to train transformer models that crushed the competition’s classic recsys task.

“Merlin provides great results right out of the box, and the flexible design lets me customize models for the specific challenge,” he said.

Riding the RAPIDS

Like his teammates, he also used RAPIDS , a set of open-source libraries for accelerating data science on GPUs.

For example, Deotte accessed code from NGC , NVIDIA’s hub for accelerated software. Called DASK XGBoost, the code helped spread a large, complex task across eight GPUs and their memory.

For his part, Titericz used a RAPIDS library called cuML to search through millions of product comparisons in seconds.

The team focused on session-based recommenders that don’t require data from multiple user visits. It’s a best practice these days when many users want to protect their privacy.

To learn more:

Watch a GTC session on building session-based recommenders with Merlin.

Take a recsys course from the NVIDIA Deep Learning Institute .

Check out the next-item prediction workflow that’s part of NVIDIA AI Enterprise , a complete software suite with the security and support businesses require.

And watch the video below.

Original URL: https://blogs.nvidia.com/blog/2023/07/12/recommendation-systems-win/