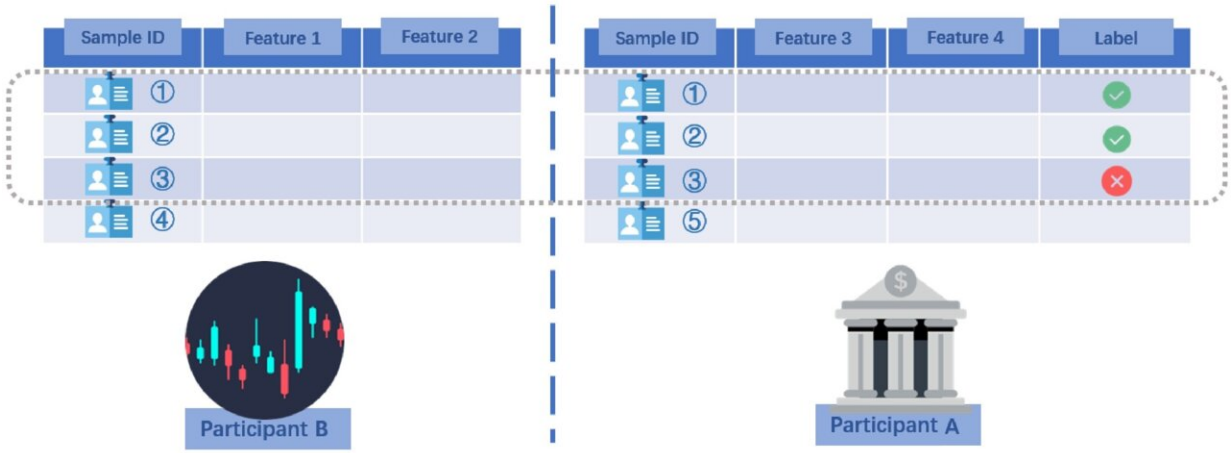
A new inference attack that could enable access to sensitive user data

March 7 2023, by Ingrid Fadelli



An example illustration of VFL. Party B is a financial company holding features 1 and 2, and party A is a bank possessing features 3 and 4. They collaborate to train a model predicting if a loan application should be approved. Credit: Morteza Varasteh.

As the use of machine learning (ML) algorithms continues to grow, computer scientists worldwide are constantly trying to identify and address ways in which these algorithms could be used maliciously or inappropriately. Due to their advanced data analysis capabilities, in fact, ML approaches have the potential to enable third parties to access private data or carry out cyberattacks quickly and effectively.

Morteza Varasteh, a researcher at the University of Essex in the U.K., has recently identified new type of inference attack that could potentially compromise confidential user data and share it with other parties. This attack, which is detailed in a paper pre-published on *arXiv*, exploits vertical federated learning (VFL), a distributed ML scenario in which two different parties possess different information about the same individuals (clients).

"This work is based on my previous collaboration with a colleague at Nokia Bell Labs, where we introduced an approach for extracting private user information in a data center, referred to as the passive party (e.g., an insurance company)," Varasteh told Tech Xplore. "The passive party collaborates with another data center, referred to as the active party (e.g., a bank), to build an ML algorithm (e.g., a credit approval algorithm for the bank)."

The key objective of the recent study by Varasteh was to show that after developing an ML model in a vertical federated learning (VFL) setting, a so-called "active party" could potentially extract confidential information of users, which is only shared with the other party involved in building the ML model. Active party could do so by utilizing their own available data in combination with other information about the ML model.

Importantly, this could be done without making an enquiry about a user from the other party. This means that, for instance, if a bank and an insurance company collaboratively develop an ML algorithm, the bank could use the model to obtain information about their own clients who are also clients of the insurance company, without obtaining their permission.

"Consider a scenario where a bank and an insurance company have many clients in common, with clients sharing some information with the bank

and some with the insurance company," Varasteh explained. "To build a more powerful credit approval model, the bank collaborates with the insurance company on the creation of a machine learning (ML) algorithm. The model is built and the bank uses it to process loan applications, including one from a client named Alex, who is also a client of the insurance company."

In the scenario outlined by Varasteh, the bank might be interested in finding out what information Alex (the hypothetical user they share with an insurance company) shared with the insurance company. This information is private, of course, so the insurance company cannot freely share it with the bank.

"To overcome this, the bank could create another ML model based on their own data to mimic the ML model built collaboratively with the insurance company," Varasteh said. "The autonomous ML model produces estimates of Alex's overall situation in the insurance company, taking into account the data shared by Alex with the bank. Once the bank has this rough insight into Alex's situation, and also using the parameters of the VFL model, they can use a set of equations to solve for Alex's private information shared only with the insurance company."

The inference attack outlined by Varasteh in his paper is relevant to all scenarios in which two parties (e.g., banks, companies, organizations, etc.) share some common users and hold these users' sensitive data. Executing these types of attacks would require an "active" party to hire developers to create autonomous ML models, a task that is now becoming easier to accomplish.

"We show that a bank (i.e., active party) can use its available data to estimate the outcome of the VFL model that was built collaboratively with an insurance company," Varasteh said.

"Once this estimate is obtained, it is possible to solve a set of mathematical equations using the parameters of the VFL model to obtain hypothetical user Alex's private information. It is worth noting that Alex's private information is not supposed to be known by anyone. Although some countermeasures additionally have been introduced in the paper to prevent this type of attack, the attack itself is still a notable part of the research results."

Varasteh's work sheds some new light on the possible malicious uses of ML models to illicitly access users' personal information. Notably, the attack and data breech breach scenario he identified had not been explored in previous literature.

In his paper, the researcher at University of Essex proposes privacy-preserving schemes (PPSs) that could protect users from this type of inference attack. These schemes are designed to distort the parameters of a VFL model that correspond to features of data held by a so-called passive party, such as the insurance company in the scenario outlined by Varasteh. By distorting these parameters to varying degrees, passive parties who collaboratively help an active party build an ML model can reduce the risk that the active party accesses their clients' sensitive data.

This recent work may inspire other researchers to assess the risks of the newly uncovered inference attack and identify similar attacks in the future. Meanwhile, Varasteh intends to examine VFL structures further, searching for potential privacy loopholes and developing algorithms that could close them with minimal harm to all involved parties.

"The primary objective of VFL is to enable the building of powerful ML models while ensuring that user privacy is preserved," Varasteh added. "However, there is a subtle dichotomy in VFL between the passive party, which is responsible for keeping user information safe, and the active party, which aims to obtain a better understanding of the VFL model and

its outcomes. Providing clarification on the model outcomes can inherently lead to ways to extract private information. Therefore, there is still much work to be done on both sides and for various scenarios in the context of VFL."

More information: Morteza Varasteh, Privacy Against Agnostic Inference Attacks in Vertical Federated Learning, *arXiv* (2023). DOI: 10.48550/arxiv.2302.05545