Contribution ID: **49** Type: **not specified**

Security evaluation of Deep Neural Networks and Gradient Boosting Decision Trees in an insurance context

Thursday, 16 May 2024 13:50 (20 minutes)

Machine learning (ML) will play an increasingly important role in many processes of insurance companies in the future [1]. However, ML models are at risk of being attacked and manipulated [2]. In this work, the robustness of Gradient Boosted Decision Tree (GBDT) models and Deep Neural Networks (DNN) in an insurance context is evaluated. It is analyzed how vulnerable each model is against label-flipping, backdoor, and adversarial example (AE) attacks. Therefore, two GBDT models and two DNNs were trained on two different tabular datasets from an insurance context. The ML tasks performed on these datasets are claim prediction (regression) and fraud detection (binary classification).
Label-flipping attacks do not present a high threat in the scenarios of this work, as the obstacles to a successful attack are particularly high in relation to the potential gain for an adversary. Nevertheless, a small fraction of flipped labels can reduce the general performance of the models drastically. in the case of backdoor attacks manipulated samples were added to the training data. It was shown that these attacks can be highly successful, even with just a few added samples. This indicates that a potentially large threat through these attacks exists. However, the success of backdoor attacks also heavily depends on the underlying training data set, illustrating the need for further research examining the factors that contribute to the success of this kind of attack. Lastly, a modified version of the Feature Importance Guided Attack [3] was used for the AE attacks. These attacks can also be very successful against both model types. Modifications of just one or few features can have a strong effect. The threat level of this attack depends on how easily those features can be manipulated in a real-world case. Additionally, this attack can be executed by an attacker with little knowledge about the ML based application.
The research shows that overall, DNNs and GBDT models are clearly vulnerable against different attacks. Past research in this domain mainly focused on homogenous data [4, 5]. Therefore, this work provides important implications regarding the vulnerability of ML models in a setting with tabular (insurance) data. Hence, depending on the application context potential vulnerabilities of the models need to be evaluated and mitigated.

Type of presentation

Contributed Talk

Primary author: KÜHLEM, Robin (Debeka)

Co-authors: Prof. MAUTHE, Andreas (University of Koblenz); Dr OTTEN, Daniel (Debeka); Dr LUDWIG, Daniel (Debeka); ROSENBAUM, Alexander (University of Koblenz); Prof. HUDDE, Anselm (Debeka and Koblenz University of Applied Sciences)

Presenter: KÜHLEM, Robin (Debeka)

Session Classification: Contributed session

Track Classification: Spring Meeting