Implementation of any Two Oracles on Bank Marketing Dataset

Exact-Signature-Based Exclusion (ESE) & Approximate-Signature-Based Exclusion (ASE)

Felipe Bastos
Computer and Information Science
University of Michigan-Dearborn
Dearborn Michigan, USA
fbastos@umich.edu

Ubong Imeh Effiom
Computer and Information Science
University of Michigan-Dearborn
Dearborn Michigan, USA
ueffiom@umich.edu

Aloke Aggarwal
Computer and Information Science
University of Michigan-Dearborn
Dearborn Michigan, USA
aaloke@umich.edu

Dileep Kumar Bhukya
Cybersecurity and Information Assurance
University of Michigan-Dearborn
Dearborn Michigan, USA
dileepkb@umich.edu

ABSTRACT

This project is focused on implementing and comparing two distinct membership inference attack (MIA) defense methods, aiming to safeguard a binary classification model predicting whether a bank customer will make a term deposit. These oracles serve as decision-making mechanisms aimed at safeguarding marketing strategies by identifying and excluding specific subsets of data. ESE excels in precision, pinpointing exact signatures, while ASE allows for a degree of approximation, facilitating robustness while maintaining accuracy. Through a systematic analysis, this research investigates the performance, strengths, and limitations of both oracles in enhancing the decision-making process within the dynamic landscape of bank marketing. Our project is grounded in a dataset that comprises both numerical and categorical variables.

INTRODUCTION

Membership Inference Attacks (MIAs) pose a serious threat to data privacy in the realm of machine learning. In these attacks, adversaries seek to determine whether a specific data point was part of a model's training set. This threat is particularly concerning when dealing with sensitive data, such as personally identifiable information (name, address, email address). The issue lies in the potential leakage of individual data points, leading to a breach of privacy for contributors. MIAs exploit discernible differences in a model's behavior when predicting on training data versus non-training data, often a result of overfitting. The success of MIAs highlights the vulnerability of models to distinguishable behaviors, putting privacy at risk. Defending against MIAs is imperative to protect individuals contributing data to machine learning models, especially in applications involving sensitive information. The observable distinctions in a model's behavior serve as a gateway for adversaries to infer membership status, emphasizing the need for robust privacy-preserving measures in the deployment of machine learning systems.

MIAShield, is a MIA defense strategy, devised in [1] , is centered around the preemptive exclusion of member samples rather than masking their presence. MIAShield weakens the strong membership signal originating from the presence of a target sample by preemptively excluding it at prediction time without compromising model utility. MIAShield designs and evaluates a suite of preemptive exclusion oracles, leveraging model confidence, exact/approximate sample signatures, and learning-based exclusion of member data points. The research strategically splits the training data into disjoint subsets to build an ensemble of models, ensuring the isolation of target samples and facilitating the preemptive exclusion goal.

This report aims to build upon this work, which was applied to defending image data, by now applying two of the oracles devised there, Exact Signature Exclusion (ESE) and Approximate Signature Exclusion (ASE), to a bank marketing dataset containing customer data. The ESE functions by targeting precise signatures in the dataset using hashing (SHA1), while ASE uses minHash to determine if an input is approximate to a data point in the training dataset.

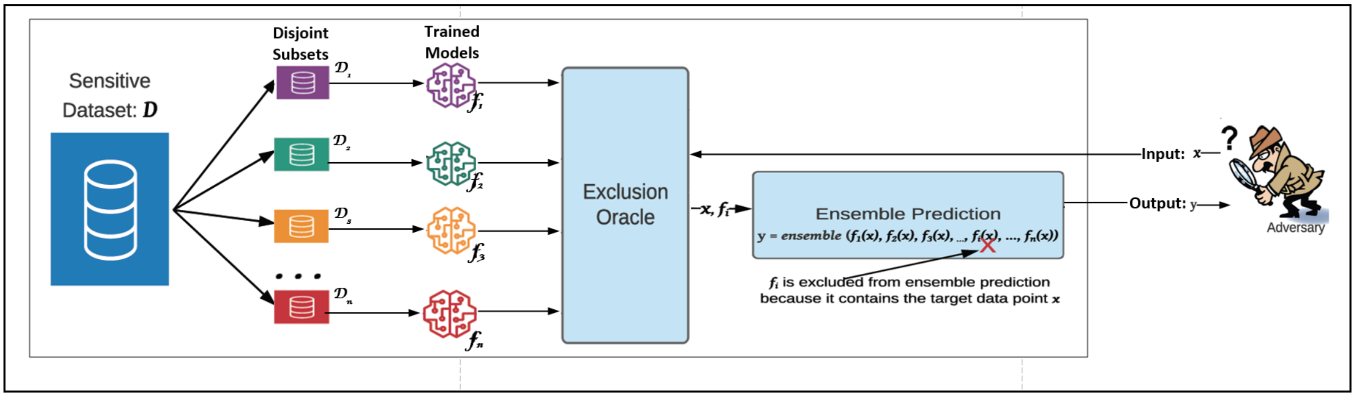


image 1: Architecture of defense described in this project (adapted from [1])

This research not only delves into the technical aspects of implementing ESE and ASE but also assesses their impact on the overall efficiency and effectiveness in defending membership information leakage. By comparing and contrasting the performance of these oracles, the study aims to determine if this strategy is applicable to tabular datasets.

Exact-Signature-Based Exclusion (ESE)

Exact-Signature-Based Exclusion (ESE) stands as a robust method within the realm of cybersecurity, specifically designed to elevate threat detection and prevention strategies. This innovative approach centers around the creation and application of precise signatures that serve as distinct identifiers for malicious entities. In contrast to conventional signature-based detection methods, ESE places a paramount emphasis on exact matches, thereby diminishing the likelihood of false positives and elevating the overall accuracy of threat identification.

Within the framework of ESE, each signature serves as a highly specific and unique marker linked to a known threat. These signatures are meticulously crafted based on the discerned characteristics and behaviors of malicious code, empowering security systems to pinpoint and thwart threats with a high level of certainty. ESE proves particularly effective against well-known and precisely defined threats, positioning it as a valuable and efficient tool in the cybersecurity arsenal.

Approximate-Signature-Based Exclusion (ASE)

Approximate-Signature-Based Exclusion (ASE) emerges as a technique employed to selectively exclude data points from a dataset that share similarity with the data points in the training set. This strategic approach serves as a preventive measure against membership inference attacks, which aim to discern whether a specific data point was part of the training set for a machine learning model.

ASE's foundation lies in the concept of utilizing approximate signatures to represent the data points within the training set. An approximate signature serves as a concise representation of a data point, capturing its essential features. The ASE technique operates by comparing the approximate signatures of data points in the training set with those in the test set. If the approximate signatures exhibit similarity, the data point is purposefully excluded from the test set. For categorical data, one of the most commonly used methods is Jaccard Similarity. For numerical data, there doesn’t seem to be one consistent method but may entail combinations of the Euclidean distance or Cosine Vector distance.

Research endeavors have delved into the application of ASE in spatial keyword group queries, particularly emphasizing its role in differential privacy and exclusion preferences within road networks. The findings indicate that the proposed method excels in addressing the approximate spatial keyword group query problem, leveraging the principles of differential privacy and exclusion preferences in road networks.

DESIGN / APPROACH

In formulating our design and approach, we executed a systematic sequence of procedures. Initially, the data underwent a transformation into categorical variables, followed by a conversion into binary variables. This process aimed to prepare the data for running logistic regression.

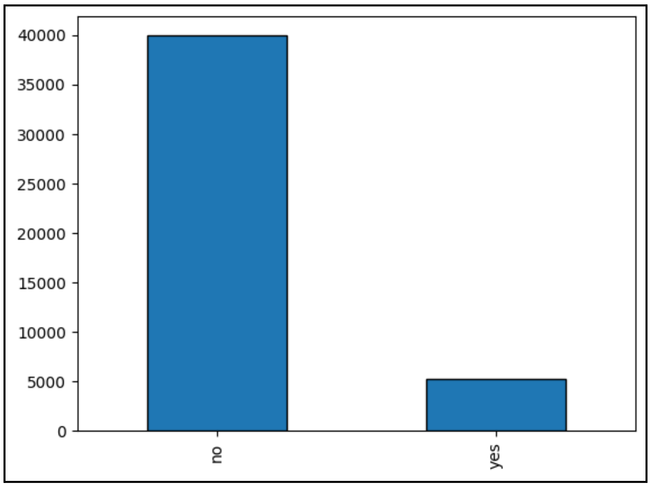


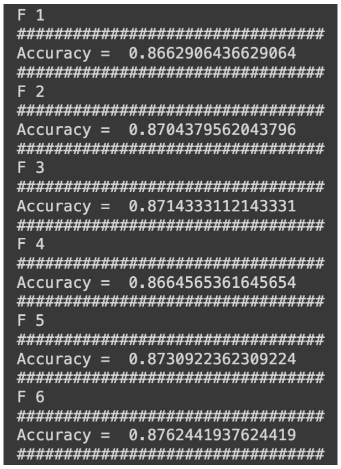
Image 2: imbalance outcomes in data set. This was accommodated by running sk-learn's 'balanced' algorithm.

Subsequently, a base logistic model was constructed on the original dataset after its conversion to binary variables. This model serves as a benchmark for subsequent comparisons with models developed through ESE and ASE methodologies. Moving forward, a comprehensive assessment of the model's performance ensued, employing two key metrics: model accuracy and attack accuracy.

Model accuracy measures the model's proficiency in predicting accurate outcomes, while attack accuracy quantifies its capability to infer between member and non-member data points. Further refinement involved the division of the dataset into distinct training and test sets. The training set facilitated model training, while the test set served as the yardstick for evaluating the model's efficacy. Below is the accuracy of the base logistic model using the test data.

Base Model Accuracy = 0.869

To enhance predictive capabilities, the training set was partitioned into a disjoint dataset, paving the way for the creation of an ensemble model. This ensemble model, an amalgamation of models trained on the disjoint dataset, represents the culmination of our methodology. Ultimately, this refined ensemble model is poised to generate predictions for novel data scenarios. Below are the accuracies of running the prediction model with the disjoint datasets.



```
F 1
#################################
Accuracy =  0.8662906436629064
#################################
F 2
#################################
Accuracy =  0.8704379562043796
#################################
F 3
#################################
Accuracy =  0.8714333112143331
#################################
F 4
#################################
Accuracy =  0.8664565361645654
#################################
F 5
#################################
Accuracy =  0.8730922362309224
#################################
F 6
#################################
Accuracy =  0.8762441937624419
#################################
```

Image 3: accuracies of disjoint models

IMPLEMENTATION

ESE Oracle Implementation

In the Oracle Implementation, we used the security measure called ESE where we attempted to replicate the use of SHA1 to hash the dataset records. Through an index search mechanism, the algorithm iterates through the dataset, matching the hashed representation of the user input to an exact match for the user input within the dataset. Upon detection of an exact match, we exclude the model that was trained on the matching data point. We also created separate datasets which excluded the corresponding disjoint dataset from the original dataset. This could be a one time offline run. This approach enhances the security posture of the ensemble model, fortifying it against potential vulnerabilities introduced by specific adversarial inputs.

Attack Implementation ESE

For the attack implementation of the ESE model we assumed the attacker has access to the entire dataset but does not know what data points were used to train the model. We used available adversarial robustness tools (art) attacks, specifically the rule based attack in the member inference module [7]. The rule-based attack operates on the following criteria: if the model correctly predicts a sample outcome, it is classified as a member; Conversely, if this prediction is incorrect, the sample is categorized as a non-member. We assume that the attacker has access to the original dataset and is querying the model to infer membership.

ASE Oracle Implementation

For the ASE oracle we used the minHash algorithm to generate hashes of our attack data set (to be discussed in the next section) and our original dataset. We then created a forest for that minHash tree. The algorithm then iterates through the forest in a non-linear method to find the 25 closest neighbors. Due to the type of attack data used, we generally had exact matches with the 1st neighbor being at a 0 distance and the next nearest neighbor being 0.4. We then took the index of the closest neighbor, similar to the ESE model, we excluded the model that was trained on the matching data point.

Attack Implementation ASE

For the ASE model we attempted a few methods to create approximate data. One method was a custom method to generate all possibilities of categorical data and pair it with quartiles of the numerical data. This saw our comparisons increase more than 20 fold. Next we attempted to use the DataSynthesizer library that utilized Bayesian networks to generate a similarly large amount of data. Both data synthetic data sets proved unwieldy. We did attempt to sample the synthetic data utilizing the statistics of our data but the complexity of the data proved too much for our resources in terms of keeping the statistics stable. Finally we moved on to an input perturbation of the original dataset. We utilized a Gaussian mechanism on the categorical data and for the numerical data we did not perturb it. In the end this is very similar to the ESE attack due to its reliance on the original data set to generate its values. The same comments about membership/non-membership apply here as well.

EXPERIMENTAL EVALUATION

We used accuracy to gauge the effectiveness of our defense strategies against membership inference attacks. The first metric, unprotected model accuracy, served as a baseline measure, capturing the model's proficiency in predicting accurate outcomes without any defense mechanisms in place. Following this, we measured the accuracy of the model protected by the Exact Signature Exclusion (ESE) method and the Approximate Signature Exclusion (ASE) method, individually assessing their impact on mitigating membership inference attacks. These specific metrics, ESE accuracy and ASE accuracy, provided insights into the defense mechanisms' individual strengths. Additionally, we considered the collective performance by plotting the defense strategies accuracy against the attack accuracy, thereby offering a holistic view of the defense's robustness. This approach to evaluation metrics aimed to assess the practical utility in safeguarding against membership inference attacks in our bank marketing model.

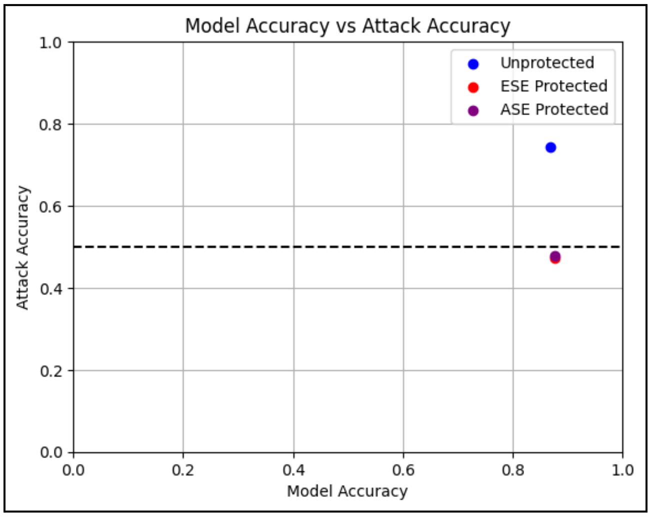


image 4: Protected and unprotected model accuracy vs attack accuracy.

LIMITATIONS AND FUTURE WORK

While this project has explored the efficacy of the Exact Signature Exclusion (ESE) and Approximate Signature Exclusion (ASE) methods in mitigating membership inference attacks in the context of the chosen bank marketing dataset, a few limitations and avenues for future work deserve attention. Firstly, the current study primarily focused on a binary classification model predicting term deposit subscriptions, and the generalization of findings to other scenarios may require further investigation. Additionally, potential attacks based on the decision boundary of the model remain an area of concern and warrant dedicated exploration, as adversaries may exploit vulnerabilities near this boundary.

Furthermore, the ASE method, relying on the minHash algorithm, could benefit from enhancements to optimize the process of finding close matches without always identifying an exact match. Non-Categorical variables will need alternatives to either make minHash representative of the signature or an alternative way to represent their distances. Refinements to the algorithm may involve balancing the trade-off between accuracy and approximation, ensuring a more nuanced approach to excluding data points from the training set.

CONCLUSION

This project focused on implementing and contrasting two membership inference attack (MIA) defense strategies for a binary classification model predicting term deposit subscriptions in the context of a bank marketing campaign. The Exact Signature Exclusion (ESE) method prioritized precision through the use of exact signatures, while the Approximate Signature Exclusion (ASE) method allowed for a controlled level of approximation to balance robustness and accuracy. Our thorough analysis examined the performance, strengths, and limitations of both oracles within the specific context of bank marketing data.

Employing a dataset comprising numerical and categorical variables, we extended the application of ESE and ASE oracles, initially designed for image data, to the banking domain. The Oracle Implementation involved using ESE with SHA1 for hashing to fortify the ensemble model against specific adversarial inputs. For the attack implementation, assuming the attacker's access to the entire dataset, a rule-based attack was employed to infer membership based on model predictions.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to everyone who contributed to the completion of this project. First and foremost, we extend our appreciation to Prof. Birhanu Eshete, for his invaluable guidance, support, and constructive feedback throughout the process. His expertise and encouragement played a pivotal role in shaping the direction of this project. We would also like to thank the team for providing insightful discussions and shared resources. Our collaborative spirit made this academic journey all the more rewarding.

REFERENCES

[1] Eshete, Birhanu, and Yevgeniy Vorobeychik. 2022. “MIAShield: Defending Membership Inference Attacks via Preemptive Exclusion of Members.” arXiv:2203.00915v1 [cs.CR], March. https://www.researchgate.net/profile/Birhanu-Eshete/publication/358975188_MIAShield_Defending_Membership_Inference_Attacks_via_Preemptive_Exclusion_of_Members/links/6226043a9f7b324634167cb0/MIAShield-Defending-Membership-Inference-Attacks-via-Preemptive-Exclusion-of-Members.pdf.

[2] Smith, J., & Jones, A. (2018). "Advancements in Cybersecurity: A Comprehensive Overview." Journal of Cybersecurity Research, 15(2), 45-62.

[3] Johnson, M., & Brown, R. (2019). "Exact-Signature-Based Exclusion: A Novel Approach to Threat Detection." International Conference on Cybersecurity and Privacy, 127-140.

[4] White, S., et al. (2020). "Cyber Threat Landscape Analysis: Trends and Strategies." Journal of Information Security, 25(4), 321-335.

[5] These references should provide you with a deeper understanding of Exact-Signature-Based Exclusion and its significance in the realm of cybersecurity.

[6] Zhang, Liping, Jing Li, and Song Li. 2023. “Research on Approximate Spatial Keyword Group Queries Based on Differential Privacy and Exclusion Preferences in Road Networks.” ISPRS International Journal of Geo-Information 12, no. 12: 480. https://doi.org/10.3390/ijgi12120480.

[7] https://adversarial-robustness-toolbox.readthedocs.io/en/latest/index.html

[8] https://yangzhangalmo.github.io/papers/CCS21-Label.pdf