Report: Membership Inference Attack on a Patient Satisfaction Prediction Model

Anthony Frank, Bradley Yong, and Catherine Zhang

1. Introduction and Motivation

Machine learning models are increasingly deployed in sensitive domains such as healthcare, where predictions draw on personal and demographic information. While such models can provide valuable insights for improving operations and patient outcomes, their use also creates potential privacy risks. In this project, we conducted a Membership Inference Attack (MIA) on a fictional hospital's patient satisfaction prediction model to illustrate one of these risks in a realistic healthcare setting.

The target model was designed to predict patient satisfaction across four categories — *Very Dissatisfied*, *Dissatisfied*, *Satisfied*, and *Very Satisfied* — using survey and demographic data, including patient age, gender, ethnicity, and distance traveled to the hospital. Although the hospital aimed to better understand patient experiences, we sought to explore how, even in well-intentioned deployments, such models can inadvertently expose private information.

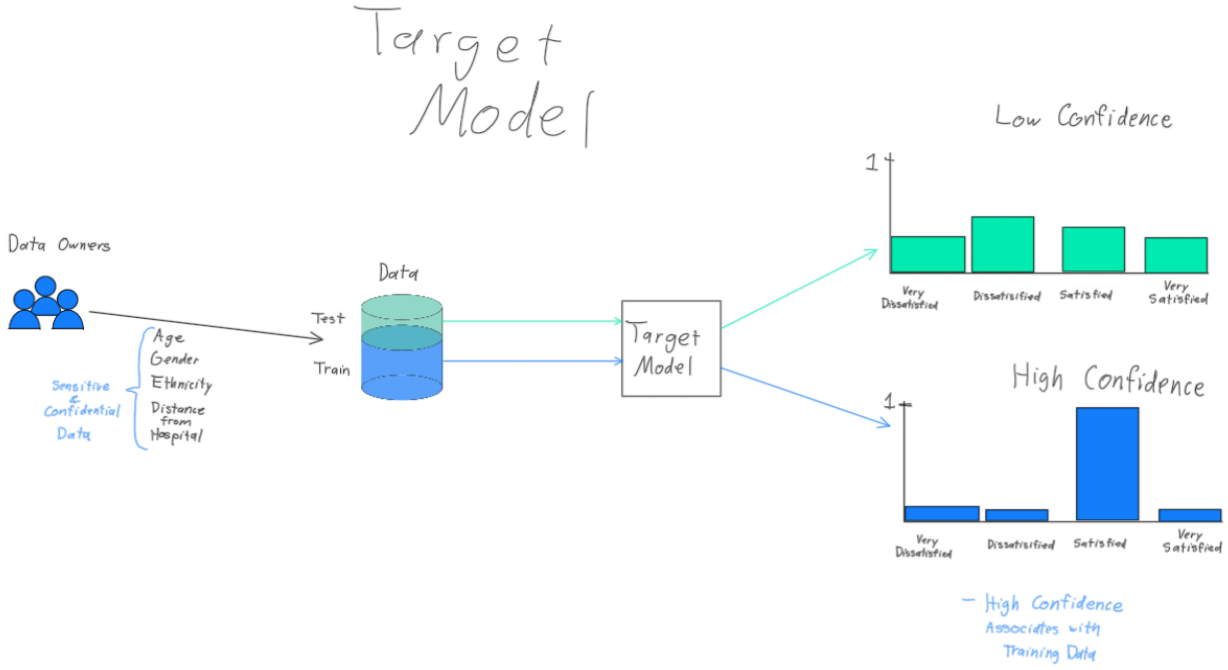
Membership Inference Attacks operate by determining whether a particular individual's record was present in the training dataset of a model. This is a serious privacy concern because, if successful, the attack can be used to reconstruct sensitive personal attributes or confirm an individual's participation in a dataset. In contexts like healthcare, this can directly violate patient confidentiality.

Previous work by Shokri et al. (2017) established that deep learning models are especially vulnerable to MIAs. The core vulnerability arises because many machine learning models — particularly over-parameterized neural networks — tend to memorize their training data and generate output probability distributions that subtly differ between data points they have seen and those they have not. Our study reproduces these findings in a healthcare-specific environment, underscoring the risks of releasing predictive APIs without robust privacy-preserving safeguards.

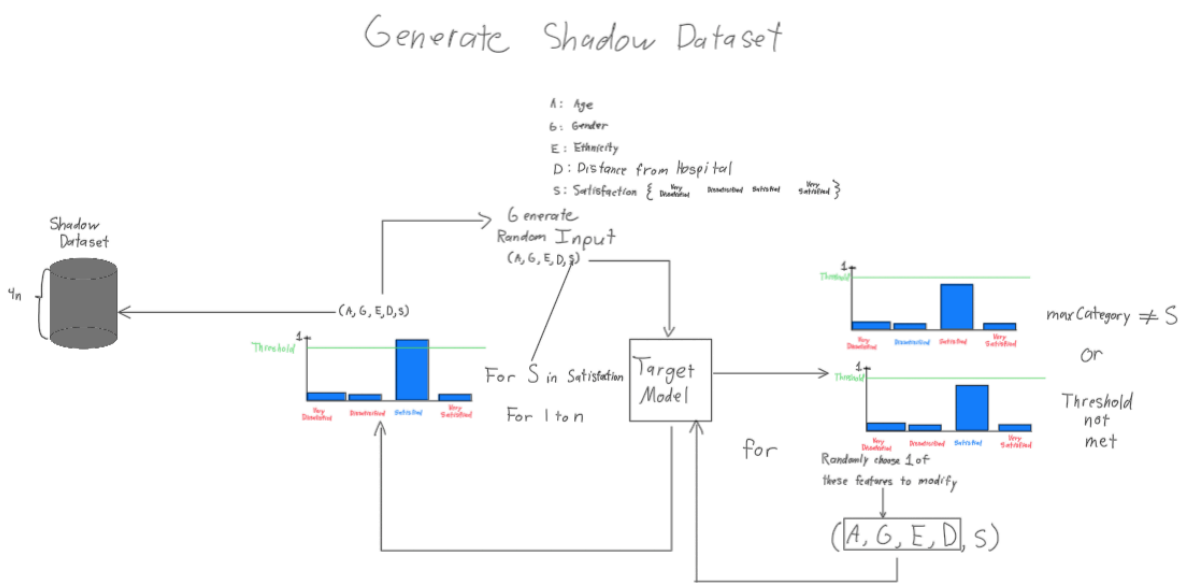
2. Results and Analysis

We began by creating a synthetic dataset simulating hospital survey data for 1,000 patients. Each record included an age between 18 and 90, gender identified as Male, Female, or

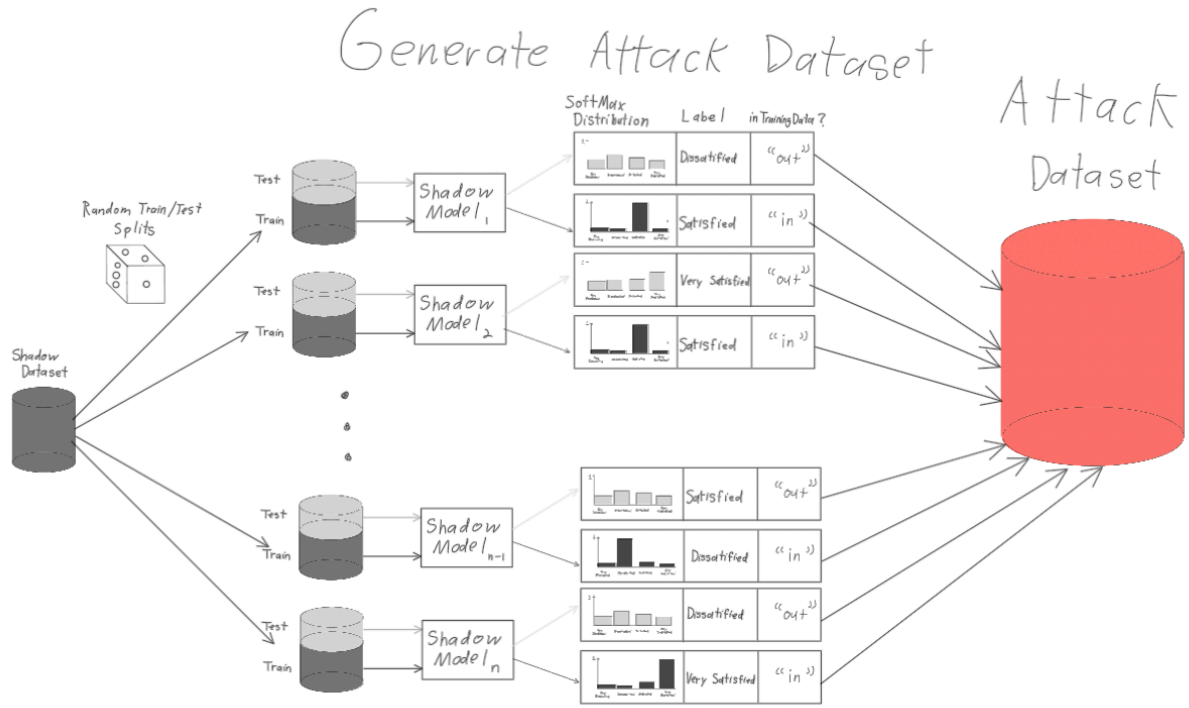
Non-binary, and ethnicity belonging to one of seven categories. We also recorded the distance each patient traveled to reach the hospital, ranging from 0.5 miles to 35 miles. The final attribute was the patient's reported satisfaction on a scale from 1 to 4. In preparing the dataset for training, all the features are categorical, so they were transformed using one-hot encoding to produce binary indicator features.



The target model was implemented as a feedforward neural network classifier with a total of 12 input features derived from the preprocessed data. The network architecture consisted of an input layer connected to a single hidden dense layer containing 12 units and employing a sigmoid activation function. The output layer comprised four neurons, one for each satisfaction category, and used a softmax activation to output class probabilities. Although the model achieved reasonable classification accuracy, the training history revealed overfitting: the training and validation loss curves began to diverge, indicating that the network had begun to memorize training examples rather than generalizing effectively to new data.

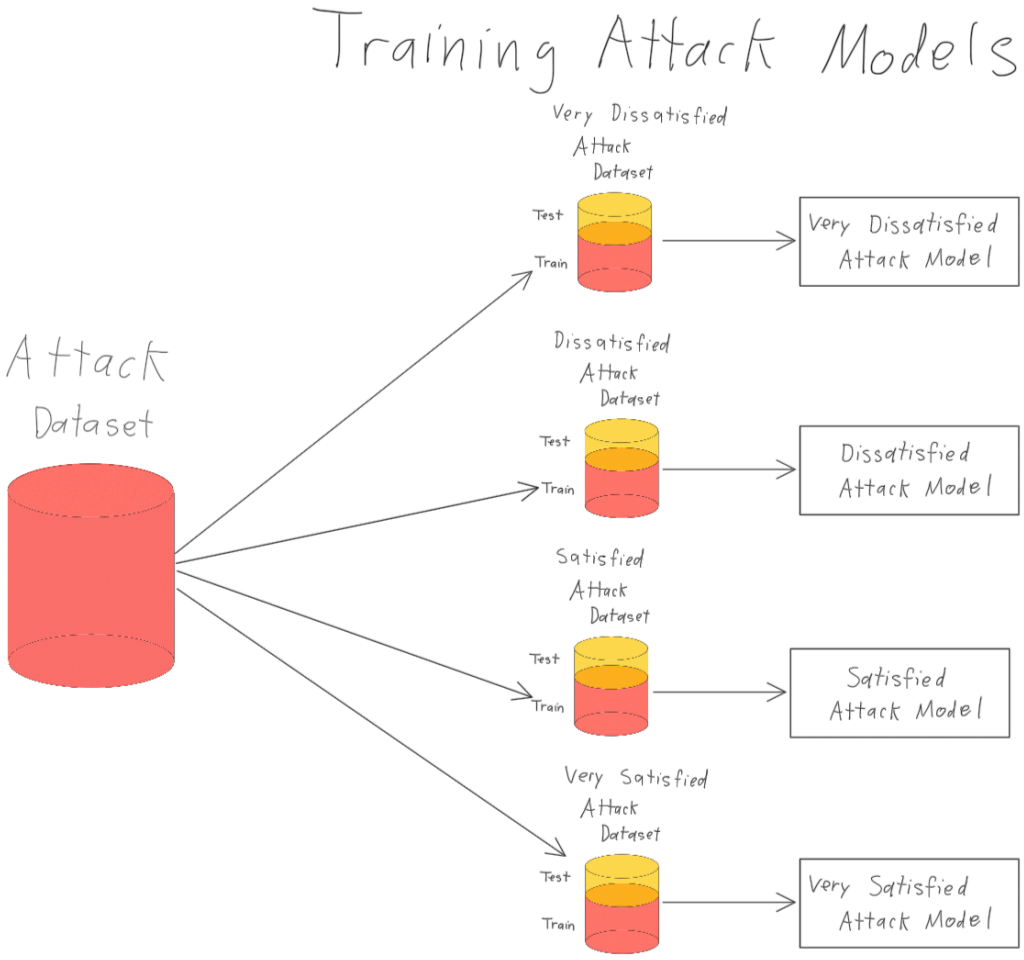


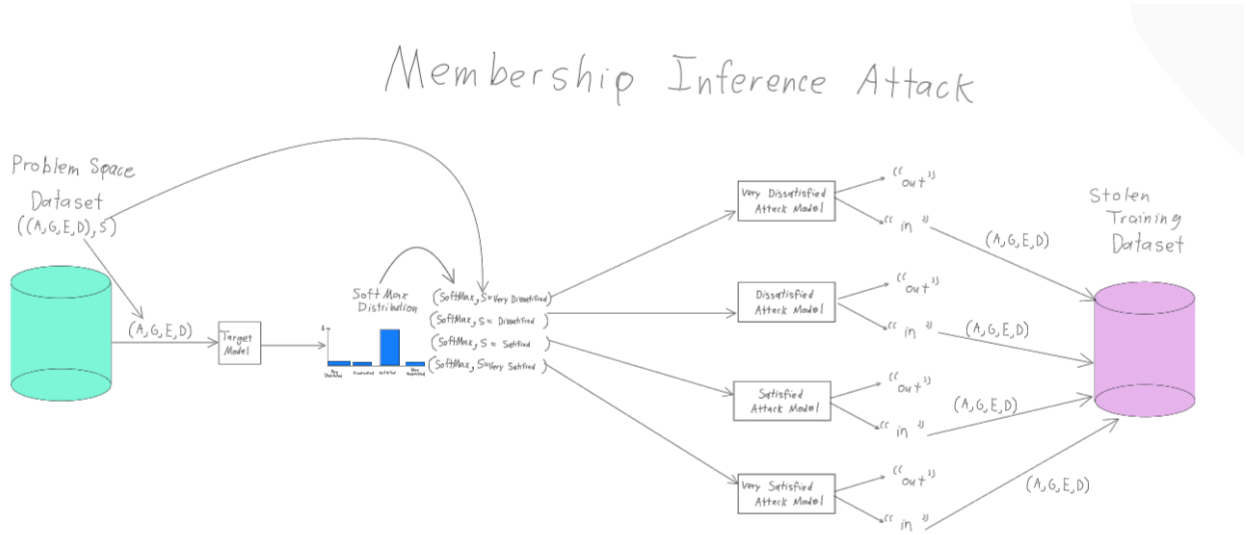
We executed the Membership Inference Attack in four distinct stages. First, we generated a “shadow” dataset to simulate the attacker's perspective when the original training data is unavailable. This was accomplished by submitting a large number of inputs to the target model and retaining only those instances where the model produced high-confidence predictions. These retained samples formed the synthetic shadow dataset.



Next, we trained multiple shadow models on this dataset, each intended to mimic the behavior of the target model. By training several variations, we captured different aspects of the decision boundaries that the original model might be using.

In the third stage, we constructed an attack dataset by recording the predictions of the shadow models and labeling each example as either “in” (meaning the sample was part of the shadow model’s training set) or “out” (meaning it came from the test set). This labeling scheme gave us a supervised dataset for learning the *in*/*out* classification task.





Finally, we trained a set of binary attack classifiers, one for each satisfaction category, to distinguish between “in” and “out” predictions. The hypothesis was that the classifiers would be able to detect subtle differences in the probability vectors produced for training samples versus unseen samples — differences that could be leveraged to infer membership in the original dataset.

```
[ ]  1  print(f"{round(stolen_training_data.shape[0]/X_train.shape[0],2)*100}% of the Model's training data was stolen!")

97.0% of the Model's training data was stolen!

[ ]  1  stolen_data = pd.merge(infered_data,clean_data,how='inner')
     2  display(stolen_data)
```

	18-25	26-45	46-65	66-95	Male	Female	Non-binary	White	Black/African_America	Asian	...	Native American/Alskan Native	Pacific_Islander	Two_or_More_Races	5-10_miles	11-20_miles	21-30_miles	Very_Dissatisfied	Dissatisfied	Satisfied	Very_Satisfied
0	1	0	0	0	0	1	0	1	0	0	...	0	0	0	0	0	1	1	0	0	0
1	1	0	0	0	0	1	0	1	0	0	...	0	0	0	0	0	1	1	0	0	0
2	1	0	0	0	0	1	0	1	0	0	...	0	0	0	0	0	1	1	0	0	0
3	1	0	0	0	0	1	0	1	0	0	...	0	0	0	0	0	1	1	0	0	0
4	1	0	0	0	0	1	0	0	1	0	...	0	0	0	0	0	1	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
865	0	0	0	1	0	0	1	0	0	0	...	0	0	1	1	0	0	0	0	0	1
866	0	0	0	1	0	0	1	0	0	0	...	0	0	1	1	0	0	0	0	0	1
867	0	0	0	1	0	0	1	0	0	0	...	0	0	1	1	0	0	0	0	0	1
868	0	0	0	1	0	0	1	0	0	0	...	0	0	1	1	0	0	0	0	0	1
869	0	0	0	1	0	0	1	0	0	0	...	0	0	1	1	0	0	0	0	0	1

870 rows × 21 columns

```
[ ]  1  print(f"{round(stolen_data.shape[0]/clean_data.shape[0],2)*100}% of the Survey data was stolen!")

87.0% of the Survey data was stolen!
```

The results showed that the attack models achieved an average accuracy of approximately 68% in determining whether a particular record was in the training dataset. While this accuracy might appear modest, it was significantly higher than the 50% baseline expected from random guessing in a binary classification setting. Precision and recall values varied somewhat between satisfaction categories, but all remained consistently above random.

Crucially, when the attack was run against the actual target model, it successfully identified specific records from the training set, as high as 97% of the target model’s training set. Among these was the survey entry of a hospital member who had provided personal demographic information. This confirmed that attributes such as age, gender, ethnicity, and travel distance could be inferred by an adversary, thereby validating the privacy threat posed by Membership Inference Attacks in healthcare machine learning applications.

3. Cost Analysis

From a computational perspective, carrying out this type of attack is relatively inexpensive with modern hardware. Training the target model — a small feedforward neural network — took roughly one to two minutes on either a CPU or GPU. Generating the shadow dataset, which involved submitting thousands of queries to the target model and filtering them, was the most time-consuming step, requiring approximately 20 to 30 minutes depending on batch size. Training the shadow models and attack classifiers was rapid, with each model completing in under two minutes.

At larger scales, where millions of API queries might be issued to a live model, the main expenses for an attacker would shift to API usage fees and the compute required for shadow dataset generation. Even then, the financial barrier remains low: using commercial cloud GPU instances, such an attack could likely be executed for under 10–20 USD. This practicality makes MIAs a credible threat vector for real-world systems.

From the defender's standpoint, implementing protective measures such as differential privacy, output perturbation, regularization techniques, or API result truncation inevitably introduces extra computational costs and sometimes degrades model utility. Nevertheless, these measures are essential in contexts like healthcare, where even minor privacy leaks can have serious ethical and legal implications.

4. Conclusions and Future Work

Our findings confirm that Membership Inference Attacks are a tangible and effective threat to deep learning models that expose detailed probability distributions via APIs. Even in a controlled, simplified healthcare simulation, an attacker was able to reconstruct sensitive information about patients from the model's outputs with accuracy well above chance. The success of such attacks stems from the fact that many models, particularly those not explicitly regularized for privacy, tend to memorize their training examples and reveal subtle but consistent differences in their responses to seen versus unseen data.

Looking forward, several avenues warrant further exploration. First, defensive strategies such as differential privacy, output perturbation, and stronger regularization should be evaluated in this context to determine how much they can reduce leakage without unduly harming model performance. Second, applying MIAs to larger and more complex real-world healthcare datasets would help quantify their effectiveness under more challenging conditions. Third, beyond membership inference, related attacks such as model inversion (reconstructing actual feature values) and attribute inference (predicting unknown features) should be investigated to understand the broader landscape of privacy vulnerabilities. Finally, this line of work should culminate in concrete deployment guidelines for hospitals and healthcare technology providers to design APIs that minimize exposure while still offering useful predictive capabilities.