Membership Inference Attack on a Patient Satisfaction Prediction Model

Anthony Frank, Bradley Yong, and Catherine Zhang

Agenda

- Introduction & Motivation
- Results & Analysis
- Cost Analysis
- Conclusions & Future Work
- Q&A

Introduction & Motivation

- Healthcare ML models use sensitive data such as age, gender, and ethnicity.
- Membership Inference Attacks (MIA) determine if an individual's data was in the training set.
- Prior work (Shokri et al., 2017) shows deep learning models are prone to MIAs.
- Goal: Demonstrate a practical MIA in a healthcare context to raise awareness about privacy risks.

Target Model Dataset

- **Size:** 1,000 synthetic patient records
- **Features:** Age (1 :18-25, 2: 26-45, 3: 46-65, 4: 66–90), Gender(1: Male, 2: Female, 3: Nonbinary), Ethnicity (7 categories), Distance to hospital(1 :5-10mi, 2: 11-20mi, 3: 21-30mi), Satisfaction score (1–4)
- **Preprocessing:** MinMax Scaling for Age and Distance, One-hot encoding for categorical data

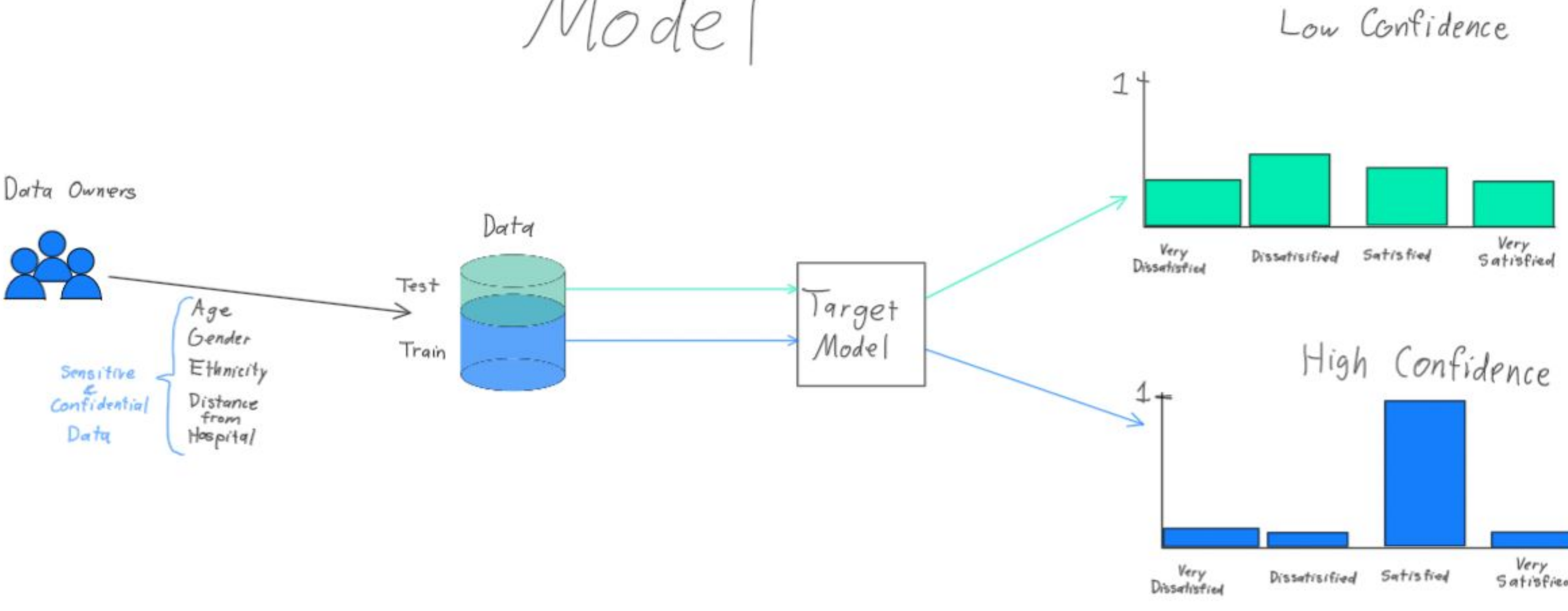
Target Model

Feedforward neural network:

- Input: 12 features
- Hidden: 12 units, sigmoid activation
- Output: 4 units, softmax activation

Achieved reasonable accuracy but showed overfitting in training curves.

Target Model

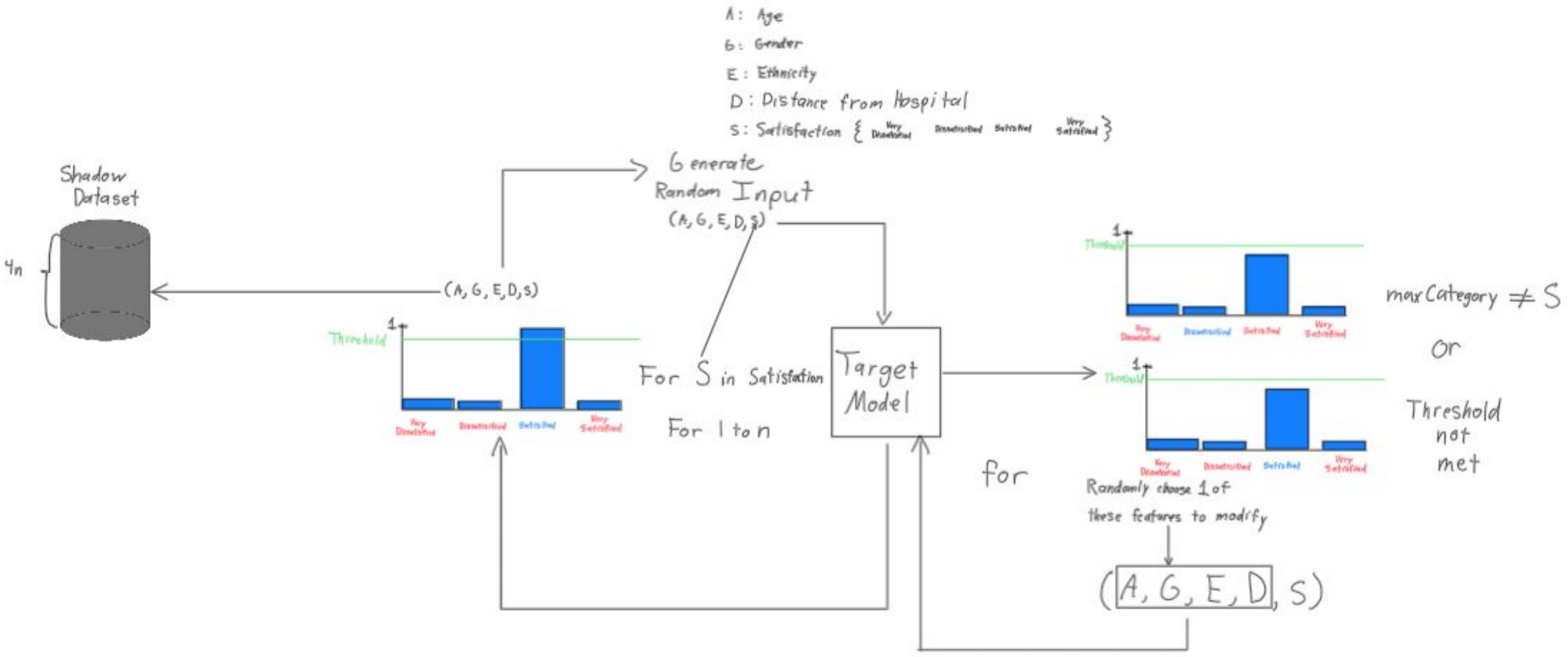


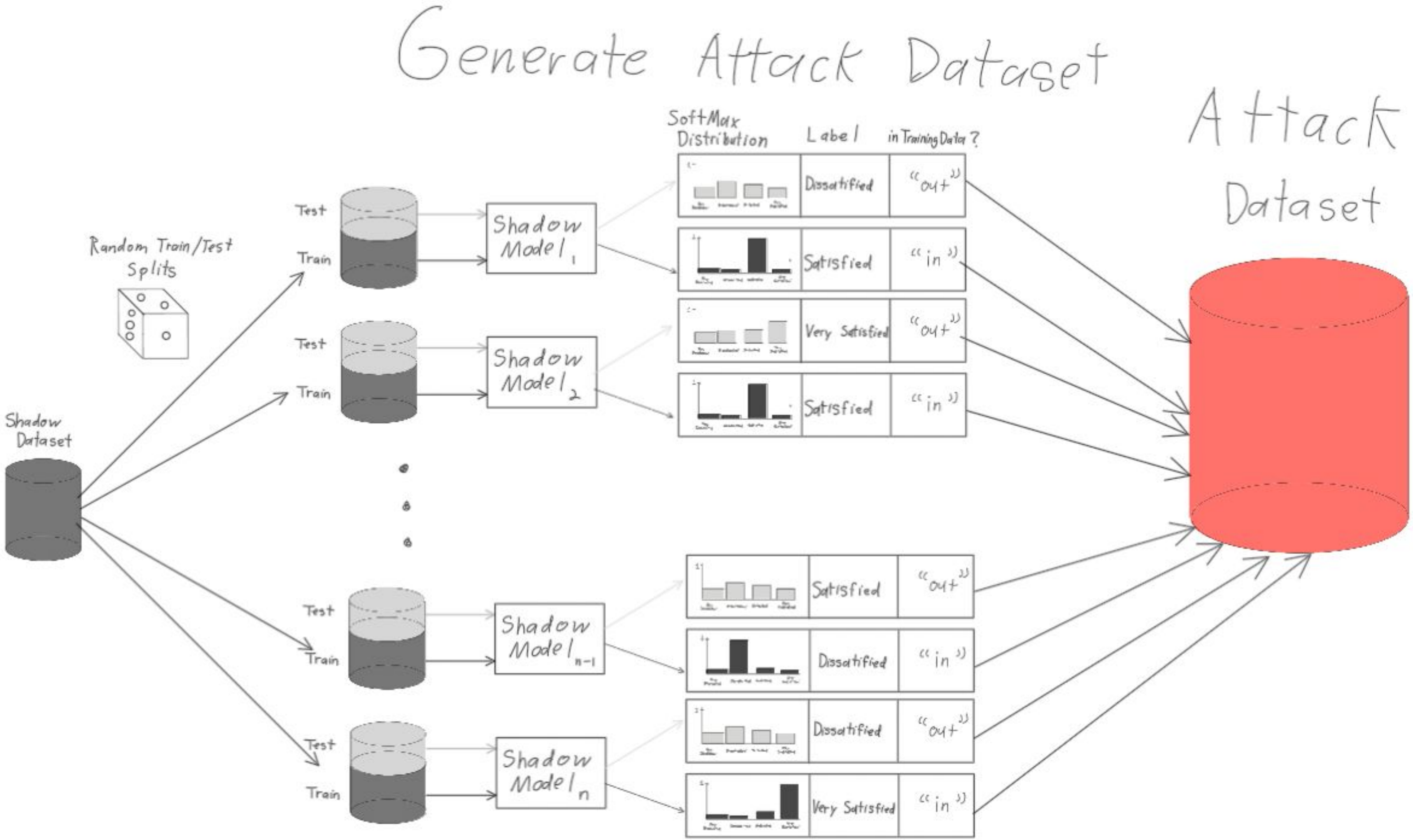
- High Confidence Associates with

Attack Methodology

1. **Shadow Dataset Generation:** Query target model, keep high-confidence samples.
2. **Shadow Models:** Train multiple models to mimic the target model's behavior.
3. **Attack Dataset:** Label predictions as *in* or *out*.
4. **Attack Models:** Binary classifiers per satisfaction category.

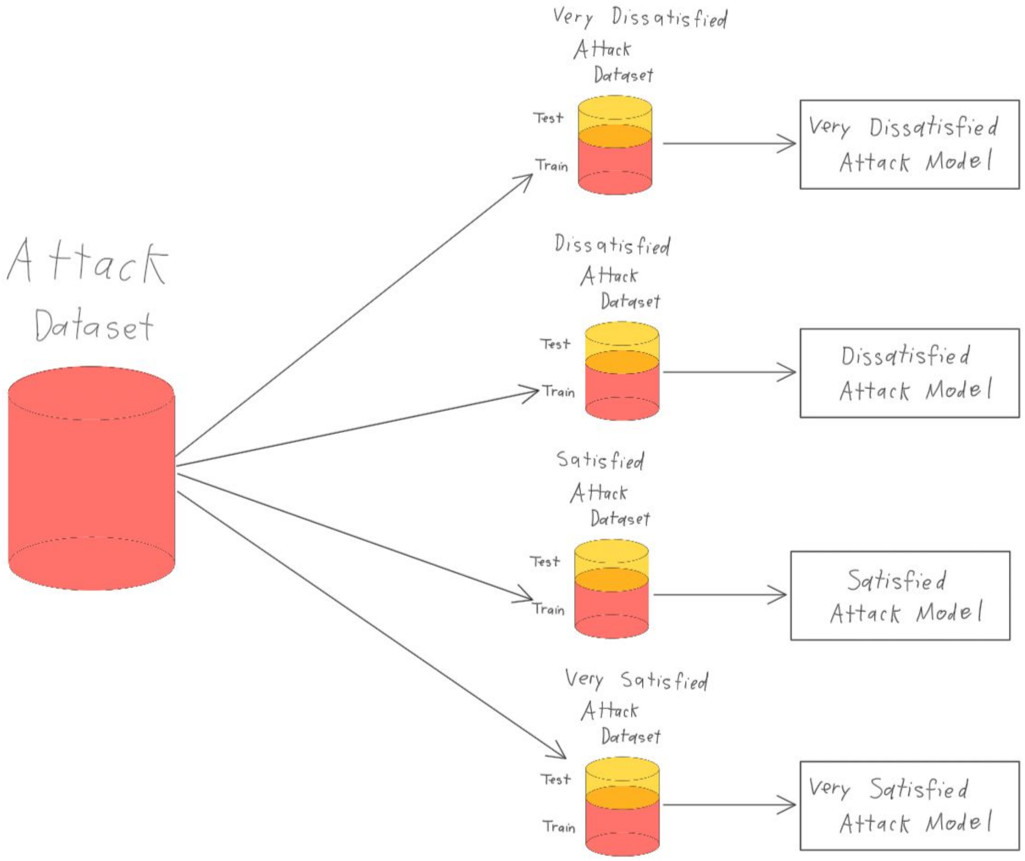
Generate Shadow Dataset



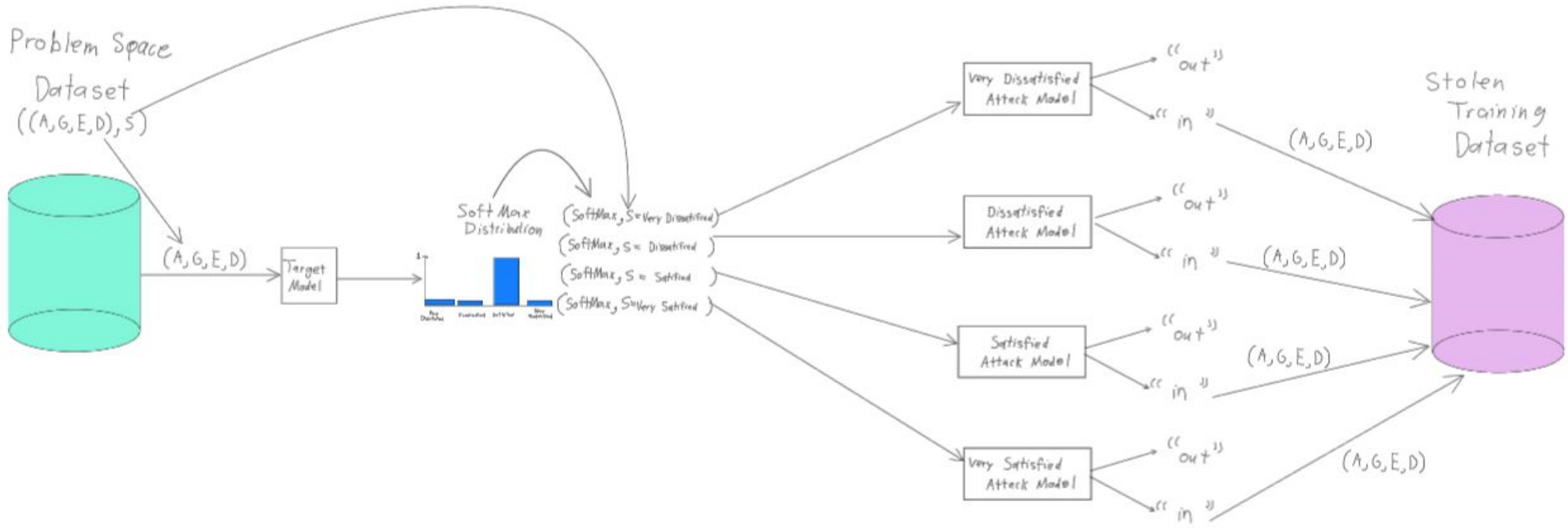


Generate Attack Dataset
Attack Dataset
Random Train/Test Splits
Shadow Dataset
Test
Train
Shadow Model 1
Shadow Model 2
Shadow Model n-1
Shadow Model n
SoftMax Distribution
Label
in Training Data ?
Dissatified
"out"
Satisfied
"in"
Very Satisfied
"out"
Satisfied
"in"
Satisfied
"out"
Dissatified
"in"
Dissatified
"out"
Very Satisfied
"in"

Training Attack Models



Membership Inference Attack



Results

- Attack accuracy: ~68% average (vs. 50% random baseline).
- Precision/recall consistently above random guessing.
- Attack recovered actual training records including sensitive attributes.
- 97% of the Model’s training data was stolen which is 87% of the Survey Data

```
[ ]  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!
```

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

```
[ ]  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!
```

Cost Analysis

Target model training: 1–2 min (CPU/GPU).

Shadow dataset generation: 20–30 min.

Shadow/attack training: < 2 min each.

Estimated attack cost at scale: <$20 USD using cloud GPUs.

Mitigation (differential privacy, output limits) adds cost but essential in healthcare.

Conclusions

MIAs are practical against deep learning models exposing probability outputs.

Even simplified healthcare simulations reveal privacy leaks.

Vulnerability arises from overfitting and confidence score differences.

Future Work

Implement & test defenses (differential privacy, output perturbation, regularization).

Apply to larger, real-world datasets.

Explore other privacy attacks (model inversion, attribute inference).

Provide API security guidelines for healthcare deployments.

Key Takeaways

Low-cost attacks can cause high-impact privacy breaches.

Sensitive domains need *privacy-first* ML design.

Proactive defenses must be integrated before deployment.

Thank you!

Questions?

Q&A