PRIVACY PRESERVING MULTI-PARTY ANALYTICS

Using GANs to Facilitate Synthetic Data Transfer with a Case Study in Credit Card Fraud Detection

Alex Langevin, Stephen Adams, Peter Beling
University of Virginia

Problem Set Up

- Many domains and industries would benefit from data sharing for predictive modeling
- Privacy, legal, competitive, or other issues may prevent data sharing between institutions
- This project investigates techniques for information transfer between parties for machine learning and model development using Generative Adversarial Networks (GANs) [1]
- The main constraint is the need for privacy – no physical transfer of, or access to underlying data by unauthorized parties
- Case study in credit card fraud detection

Goals & Objectives

- To investigate whether state of the art generative models can produce synthetic data that is realistic enough for predictive modeling
- Understand the benefits (if any) of synthetic data transfer between parties on predictive modeling performance
- Determine the feasibility of adding a differential privacy constraint to the generative model training process

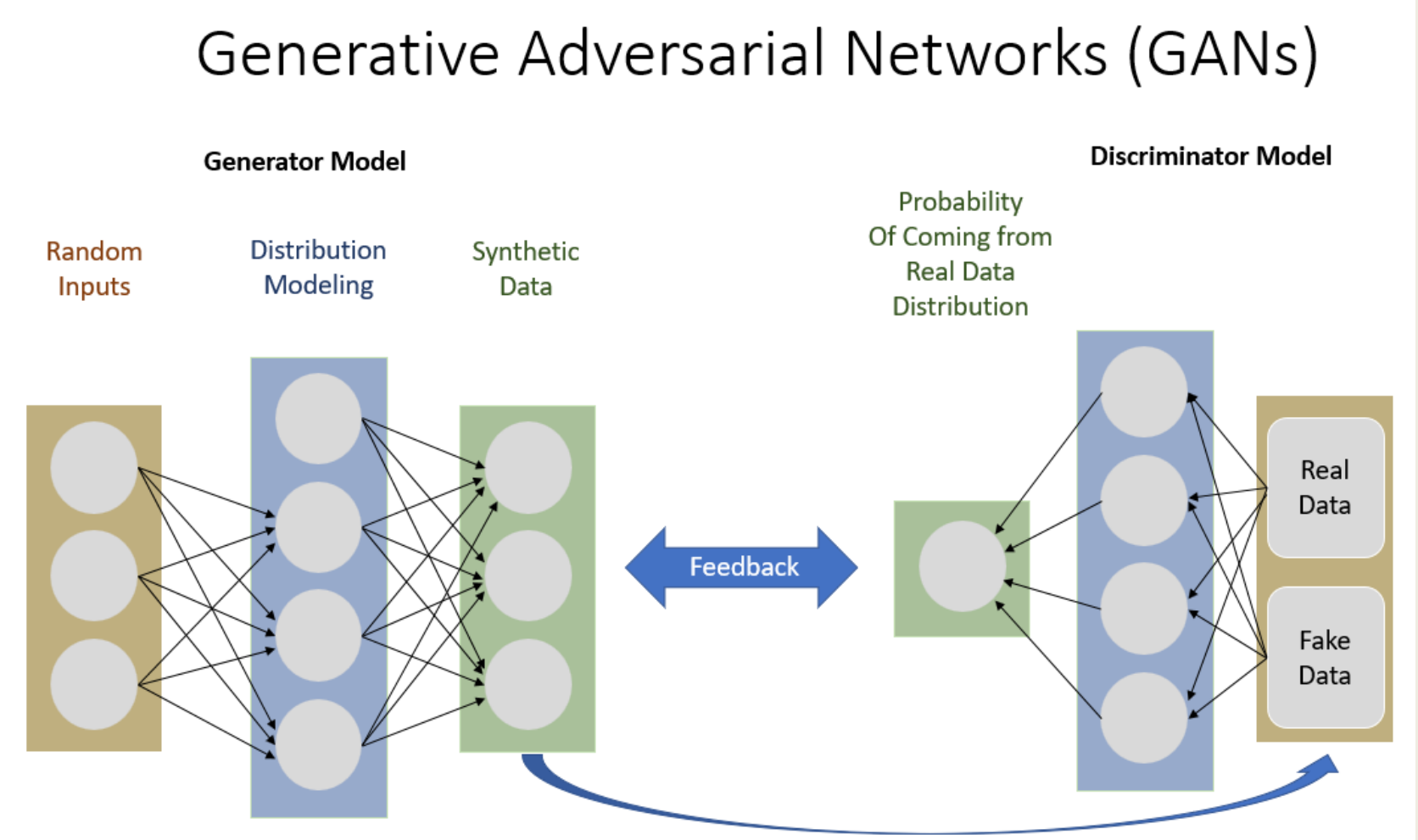
Methodology – Main Assumptions

- Two hypothetical banks in the credit card business both wish to develop fraud detection models
- Both banks also wish to limit access to transaction data by unauthorized internal and external parties
- Banks target different ends of the credit spectrum – i.e. distribution of customers are not assumed identical
- Assumption that banks are using the same feature set for modeling (to be relaxed in later work)

Methodology - Dataset

- Provided by Capital One – 77,617,325 credit card transactions over the 01/01/13 – 08/30/13 period
- Highly imbalanced – 0.14% fraud rate
- 50 original & derived features selected– 20 numeric and 30 categorical
- Categorical features: day of week, month and merchant category code
- Numeric features include e.g. credit limit, balances authorized & outstanding, timestamps, length of time since card issued, distance of purchase from home ZIP
- Derived features include e.g. proportion of card holder's transactions that fall within a given hour of day, day of week, etc.

Methodology – GANs



Methodology – GAN Refinements

- GANs are known to be fragile and difficult to train
- We adopted a more robust refinement – Wasserstein GAN (WGAN) [2] – and in particular a WGAN extension using gradient penalty in loss function (WGAN-GPs) [3]
- Mixed numeric (closed and half-closed intervals), count and categorical data:
  - *Gumbel-Softmax trick to handle categorical data*
  - *Log and logit transformations for numeric and count data with noise adjustments for boundary values*
- Also experimented with handling bounded data through Generator activation functions

Methodology – Differential Privacy

- Differential privacy [4] is a framework for placing upper bound on worst-case amount of privacy loss by adding carefully tuned amount of random noise to database query responses
- If D_1 and D_2 are databases that differ by a single record, and R is some randomized algorithm applied to the databases (our query response) which returns some output $t \epsilon T$, then the randomized algorithm R is ε-differentially private if for all outputs t,

$$\frac{\Pr[R(D_2) = t]}{\Pr[R(D_1) = t]} \leq e^{\varepsilon}$$

- We also have (ε, δ)-differential privacy,

$$\Pr[R(D_2) = t] \leq e^{\varepsilon} \Pr[R(D_1) = t] + \delta$$

- In deep learning the noise element is applied to the gradient during backpropagation
- Currently two papers utilize (ε, δ)-differential privacy and GANs
- We attempted to apply the stricter definition of ε-differential privacy [5]

Methodology – Experimental Setup

- 2 different scenarios:
  - *Two hypothetical banks A and B with similar customer profiles*
  - *Two banks with customer profiles skewed to different ends of credit spectrum*
- Dataset ordered by account credit limit (used as proxy for credit quality)
- Data partitioned into Bank A and Bank B by over/undersampling either above or below the median credit limit (MCL)
- Four datasets labelled as *Bank/Below MCL sample ratio/Above MCL sample ratio*
  - *A/55/45 vs. B/45/55*
  - *A/80/18 vs. B/20/82*

Methodology – Training Steps

- Feature Selection:
  - *Set aside 50,000 samples before data partition with 10% fraud ratio*
  - *Features selected using xgboost algorithm [6] in R*
- Non-fraud WGAN-GP Training:
  - *After data partitions, 5% of each split set aside as final test set*
  - *Hyperparameter tuning by random search on 1.5 million non-fraud transactions from each data split*
  - *Three validation criteria used – models selected for consistency across metrics*
  - *Final model trained on remaining non-fraud data with 100K holdout for early stopping epoch selection*
- Fraud WGAN-GP Training:
  - *Due to limited sample size, attempted to 'warm-up' GAN parameters on non-fraud data then transfer to fraud transactions for continued training*
  - *Done mainly to preserve privacy budget under differential privacy*

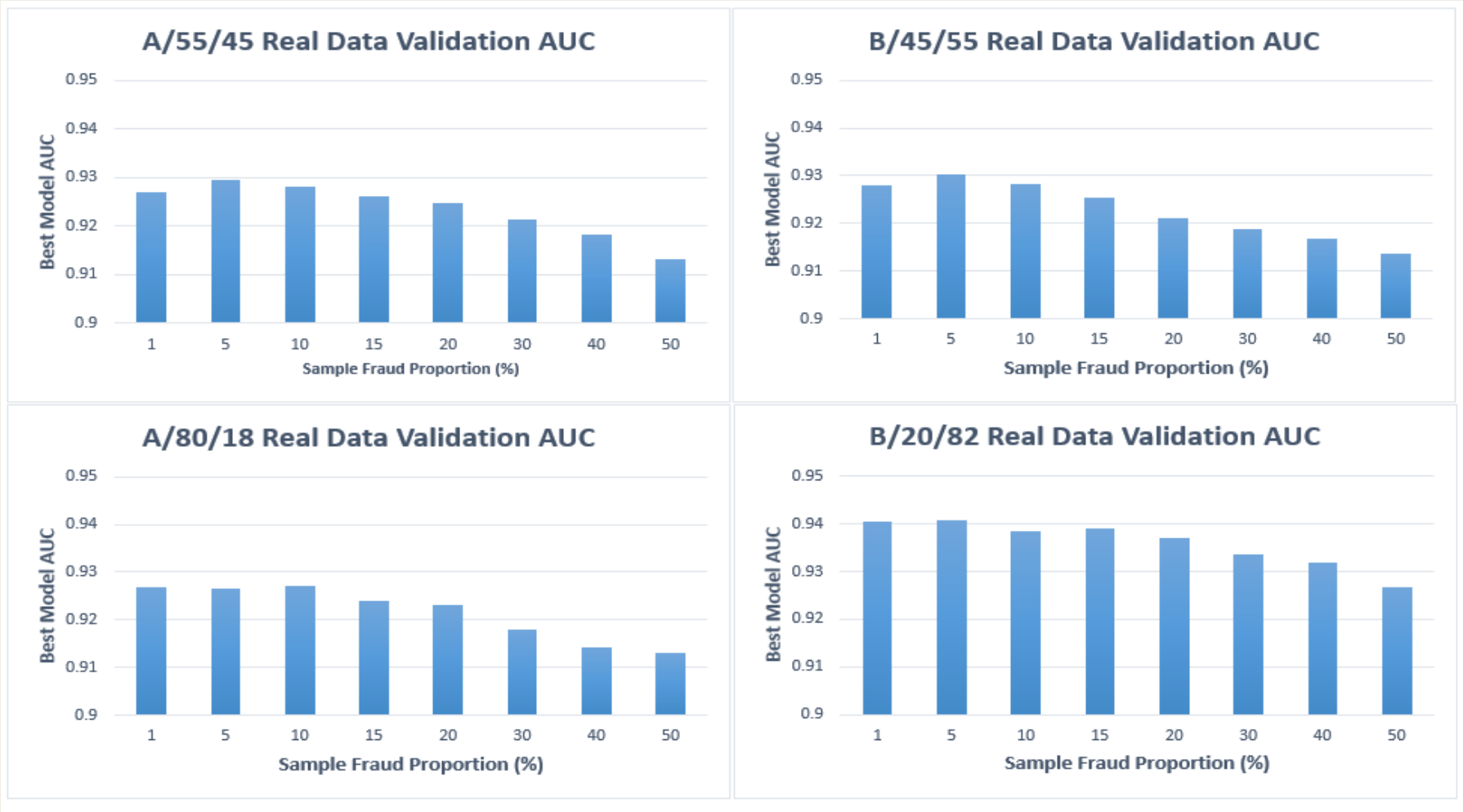
Methodology – Training Results

- After extensive hyperparameter random search, several candidate WGAN-GP candidates identified for transfer learning experiments
- In general, transferred fraud WGAN-GPs performed worse than non-fraud models but still reasonably well overall:
  - *Specific initializations for transferred models performed best*
  - *All top fraud GANs were either transferred directly, or reset final layer of critic weights*
  - *No other combination of parameters produced strong results*
- No viable ε-differentially private GAN models

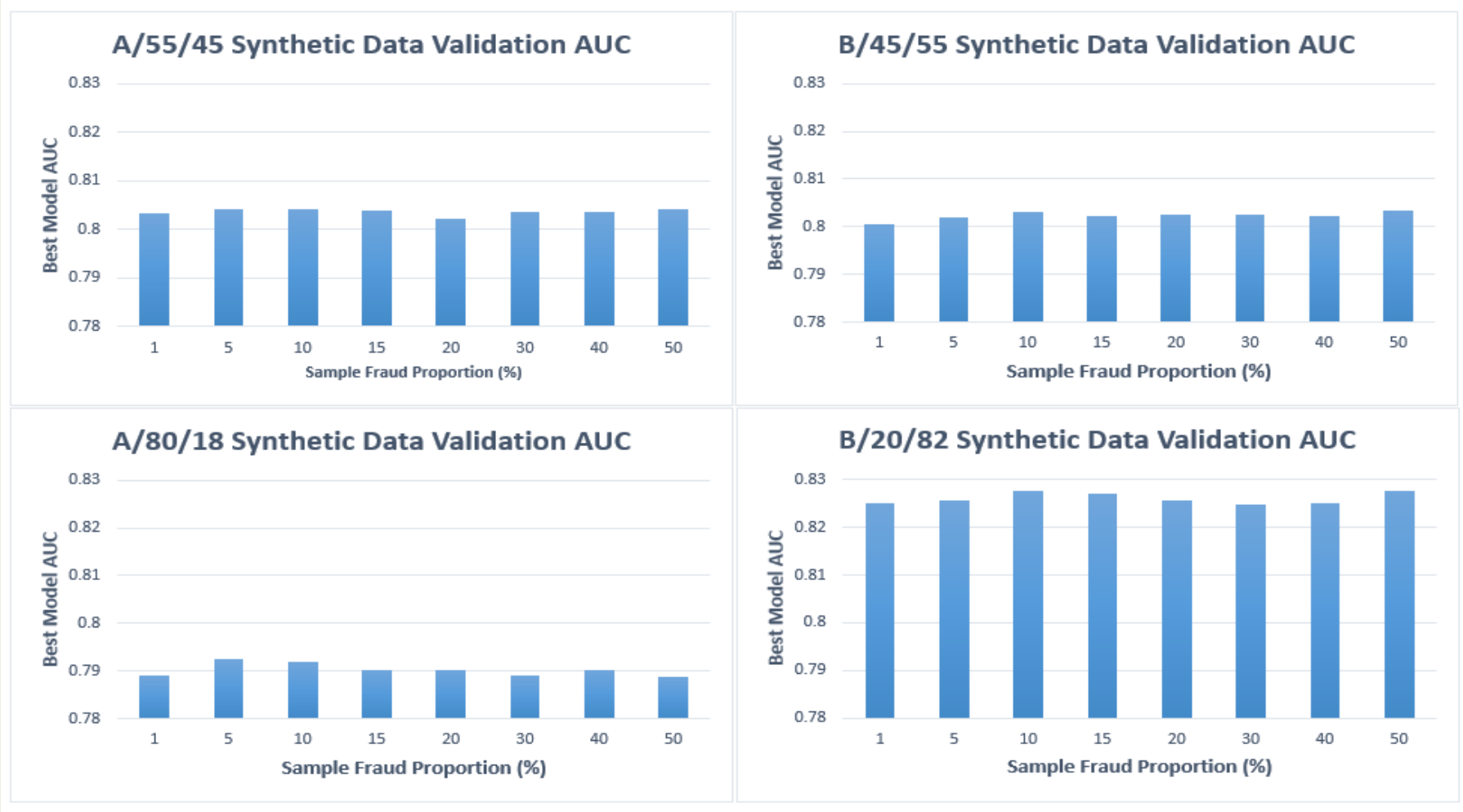
Experiments – Synthetic vs. Real Data

- For each dataset, WGAN-GPs used to generate 1 million synthetic credit card transactions at varying fraud proportions to train a fraud detection classifier
- Compared to classifiers trained on the real data, with non-fraud data downsampled to same fraud proportions
  - *Real data training sets between c.80,000 and c.4 million samples*
- Used Area Under the Curve (AUC) as performance benchmark
  - *AUC summarizes discriminatory power of a model*
  - *Compares true positive rate to false positive rate at various classification thresholds*

Results – Real Data Classifier



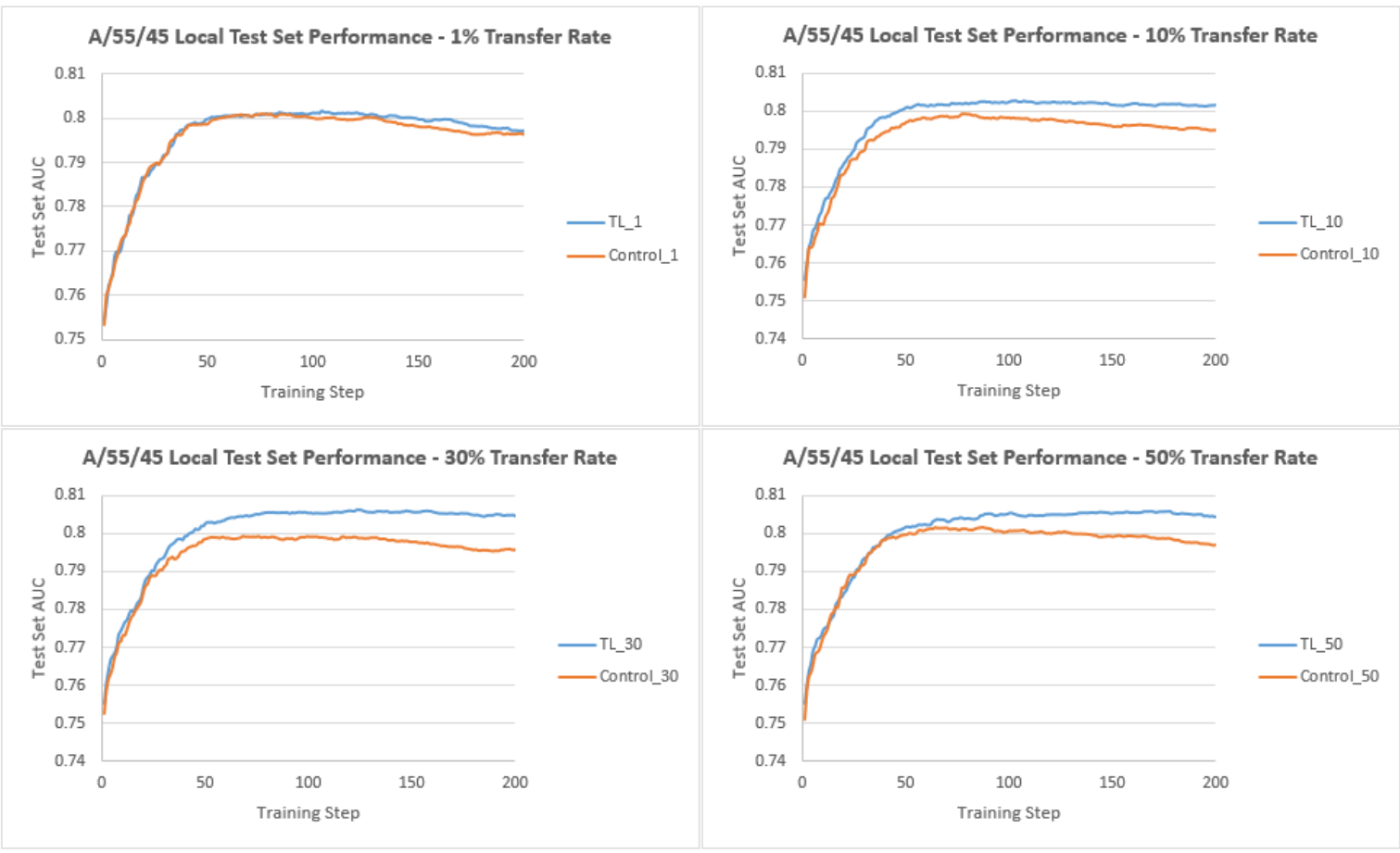
Results – Synthetic Data Classifier



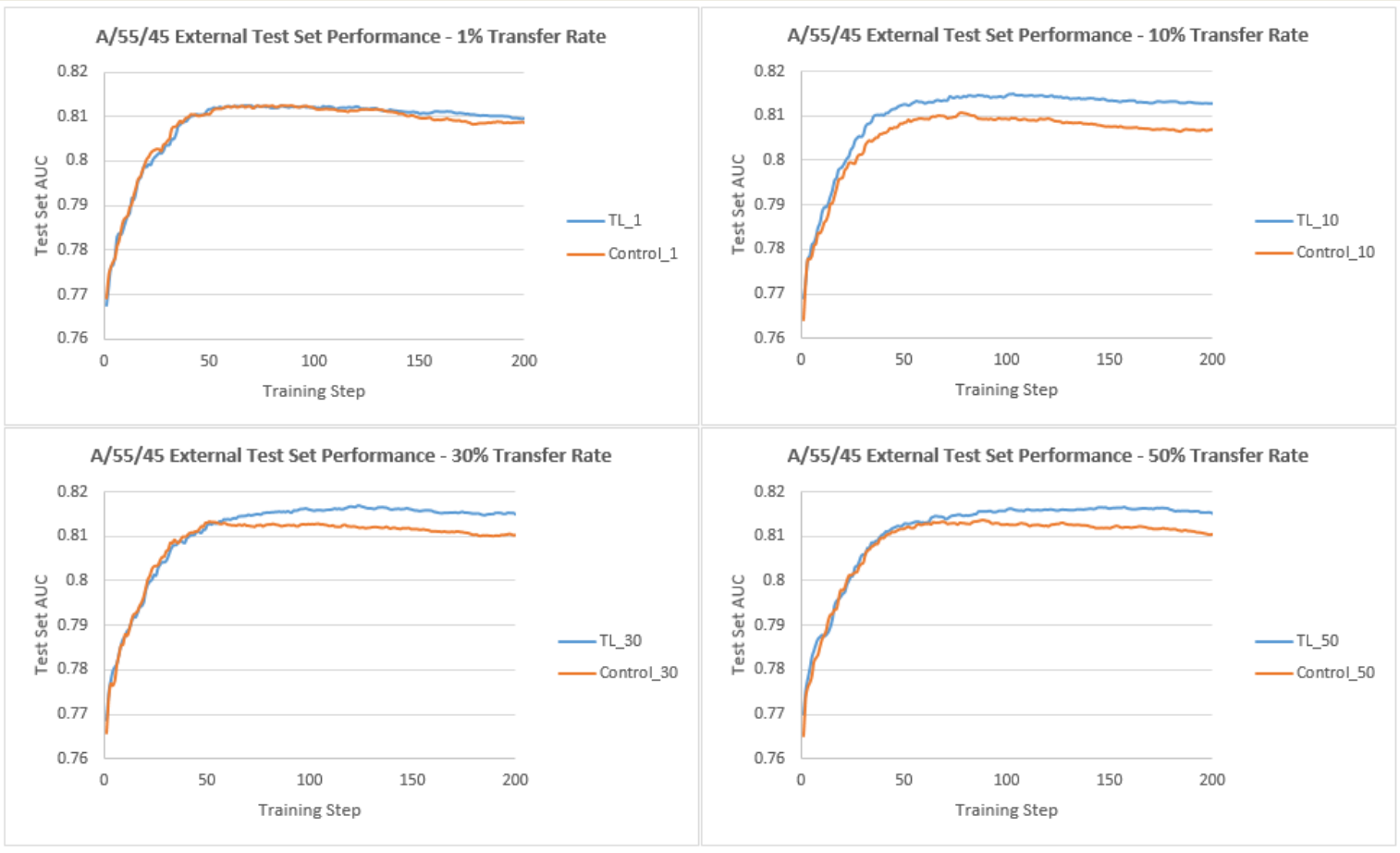
Experiments – Synthetic Data Transfer

- Fraud proportions fixed at 10%
- Each bank starts with 1 million local samples – 100,000 fraud and 900,000 non-fraud
- Data transferred to partner bank at varying proportions:
  - *1, 5, 10, 15, 20, 30, 40, 50% relative to local dataset size*
- Re-train classifier on combined data and compare performance to a control classifier trained using only local data
- Compare AUC performance on both local and partner bank test sets
  - *Partner bank test set performance considered proxy for model's ability to generalize*

Results – A/55/45 Local Performance

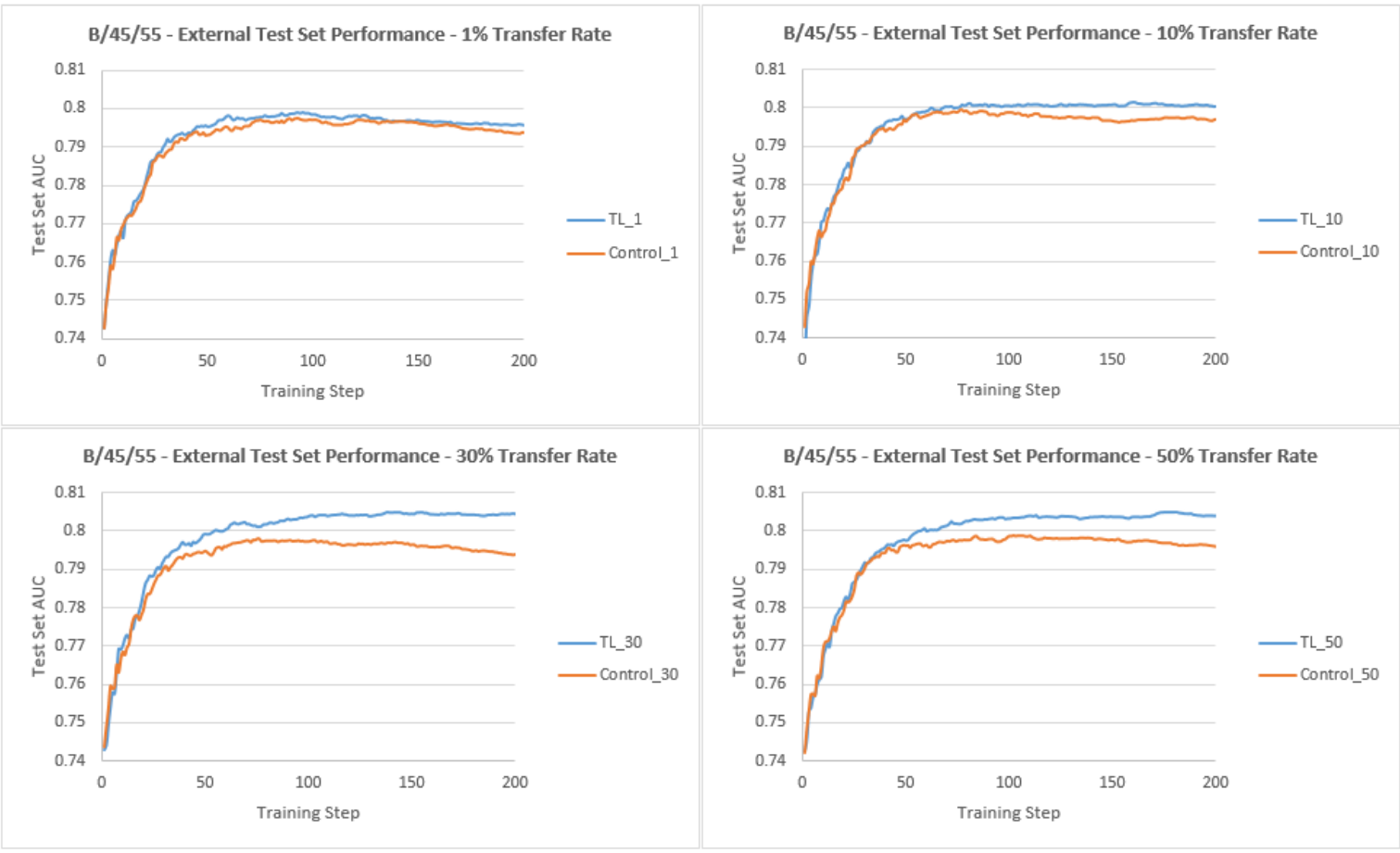


Results – A/55/45 Generalization Performance

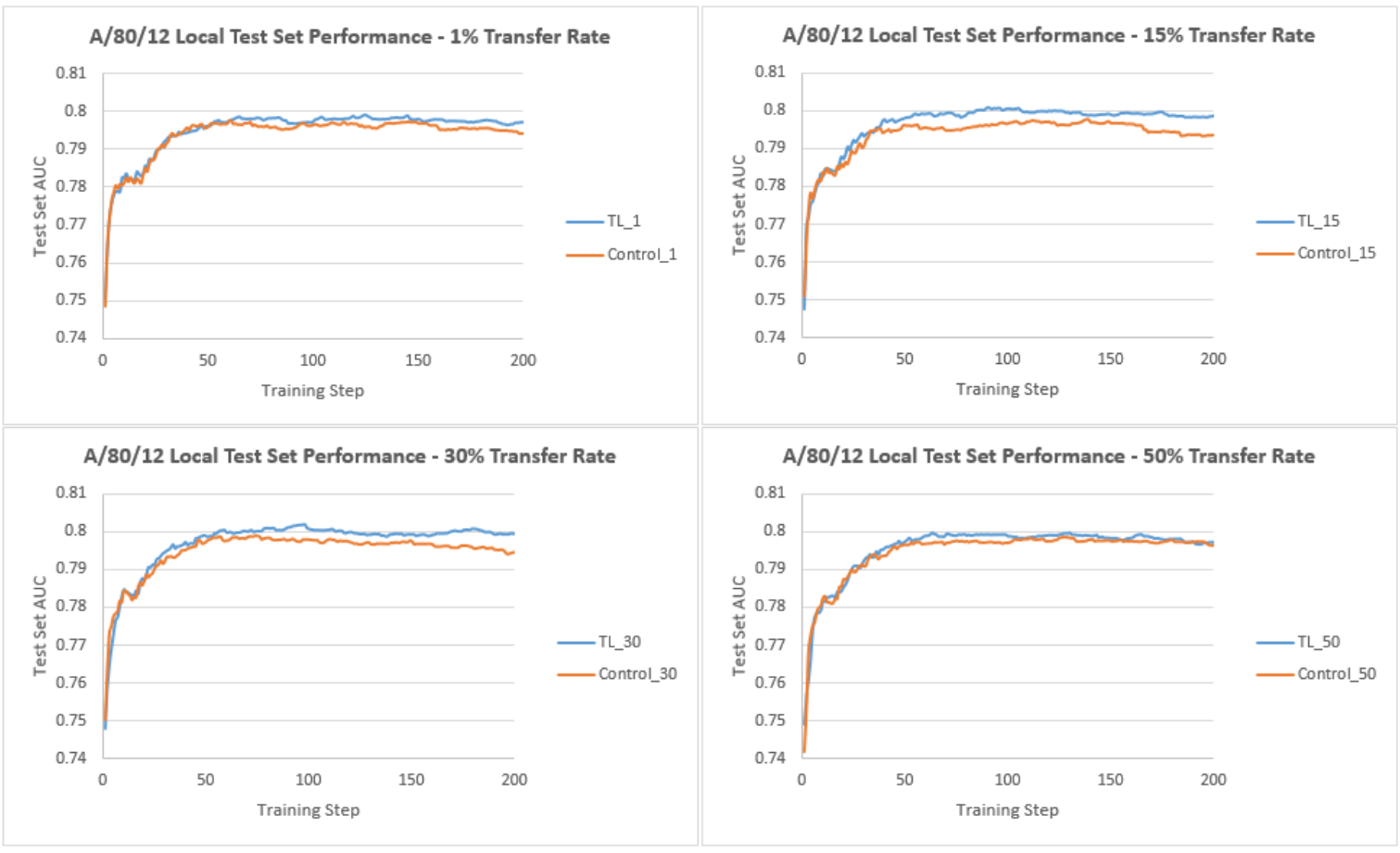


Results – B/45/55 Local Performance

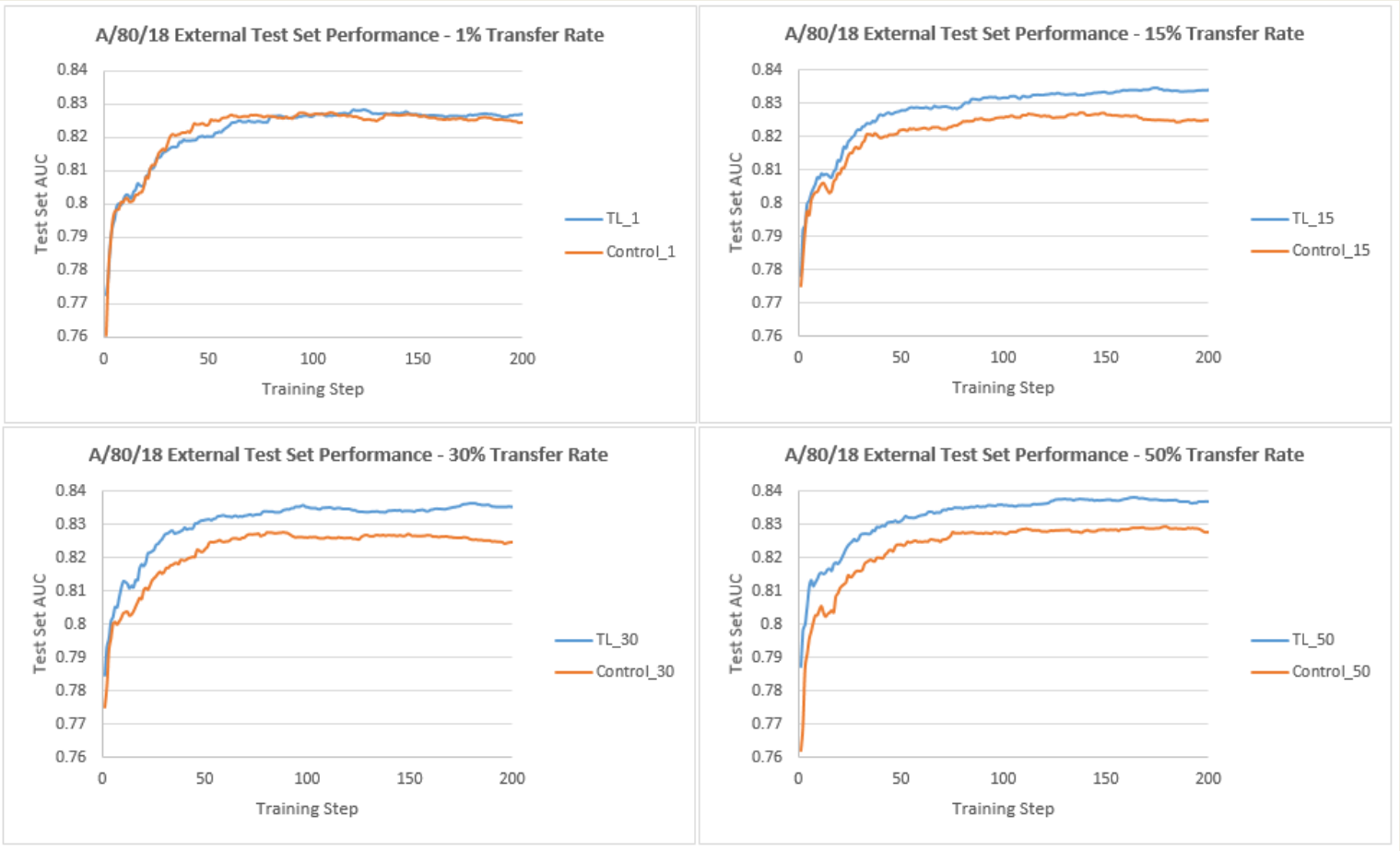
Results – B/45/55 Generalization Performance



Results – A/80/18 Local Performance

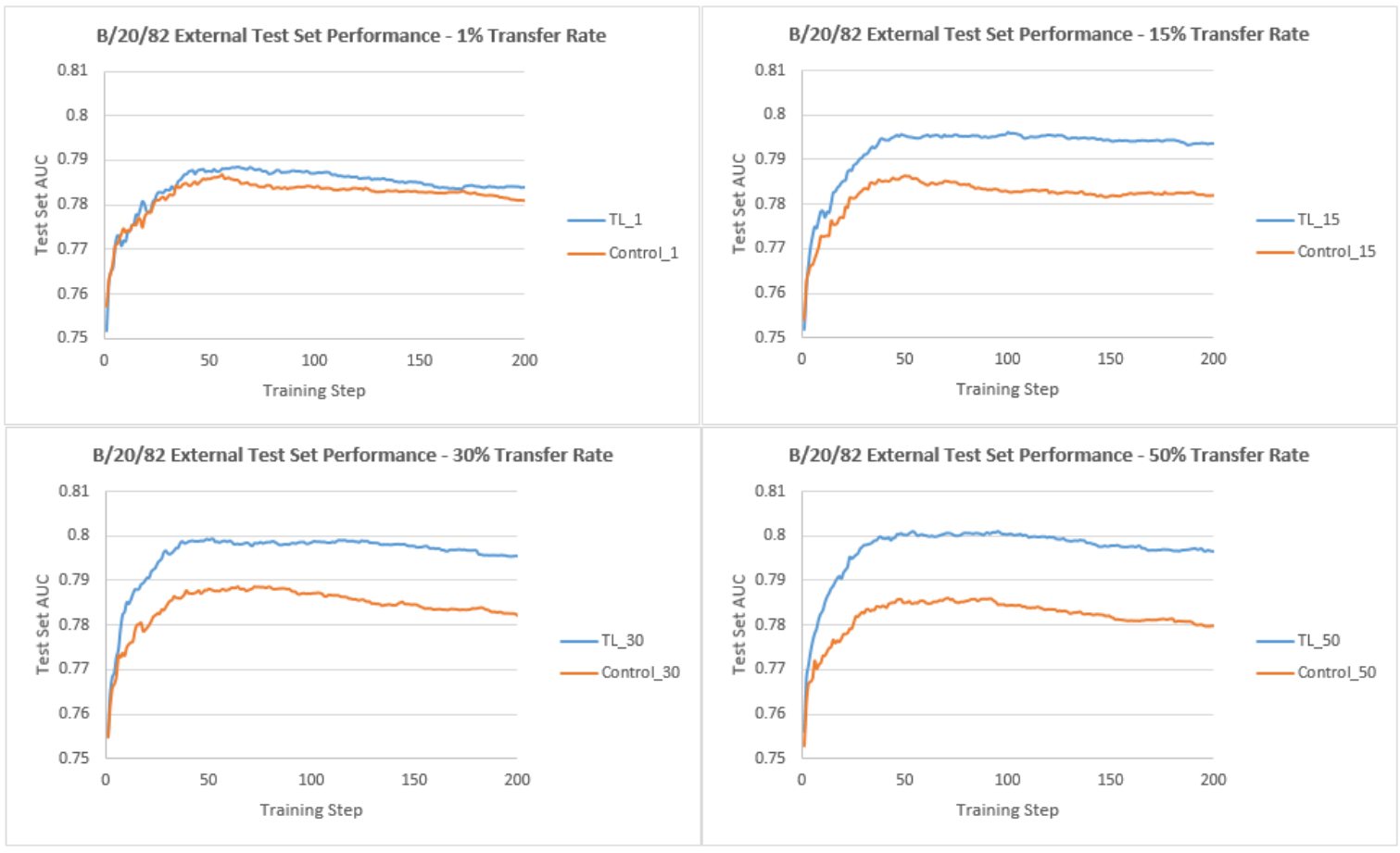


Results – A/80/18 Generalization Performance

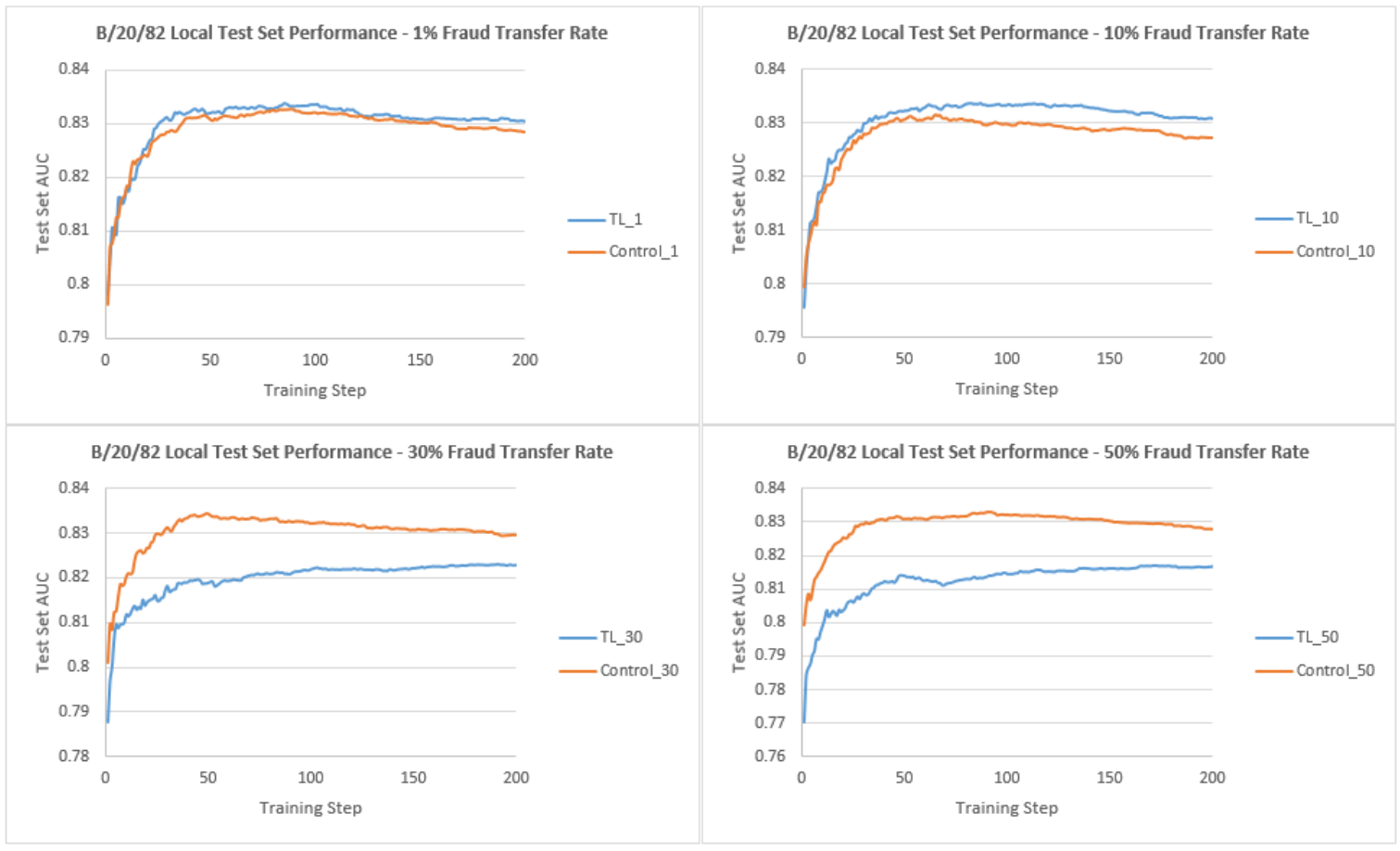


Results – B/20/82 Local Performance

Results – B/20/82 Generalization Performance



Results – B/20/82 Local Performance – Transferring Fraud Cases Only



Results – General Findings

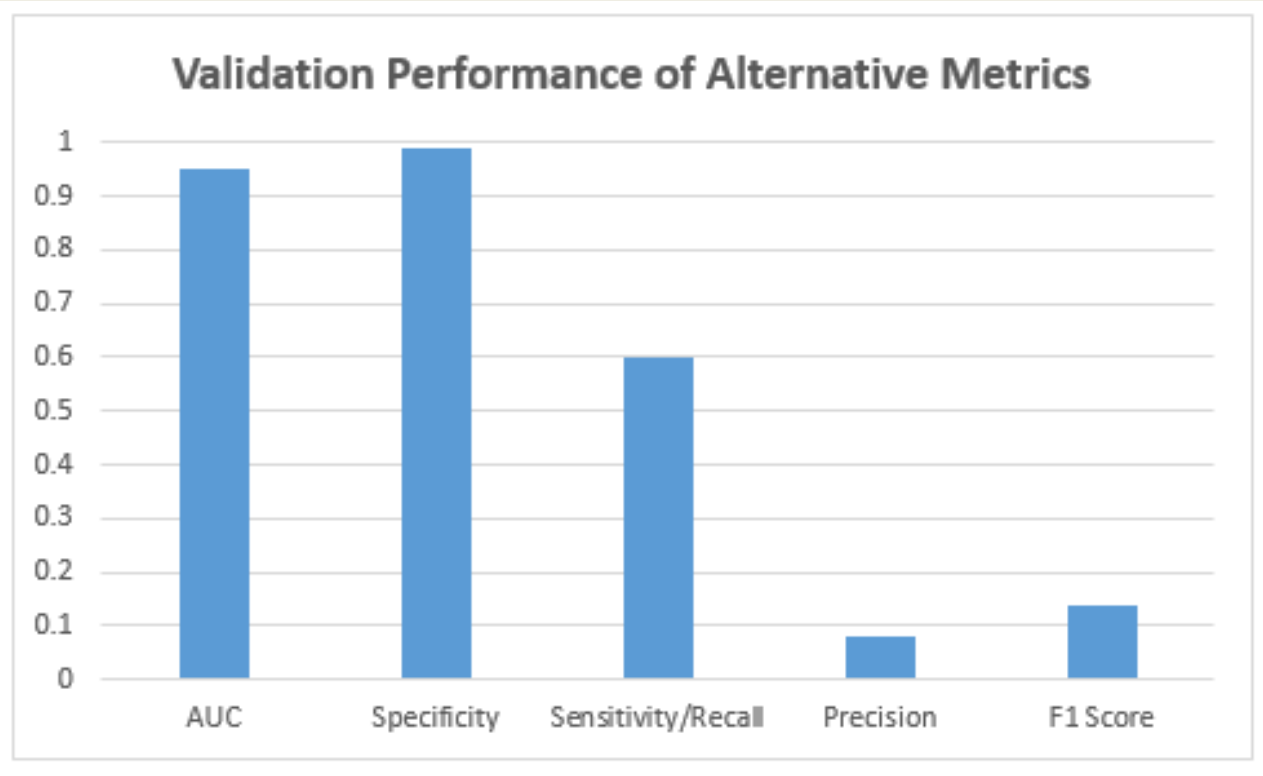
- GANs can capture enough information about the underlying customer distributions to produce classifiers that retain a significant amount of discriminatory power compared to classifiers trained on the real data
- Transferring synthetic data between banks can boost discriminatory power of models, help combat overfitting and improve generalization
- Data transfer is sensitive to differences in distributions
- Transfer benefits are asymmetric
- Important information contained in both sample classes
- Model performance appears sensitive to GAN quality and/or customer distributions

Sensitivity Analysis

- AUC may not be most appropriate metric for highly imbalanced datasets [7]
  - *AUC partly determined by False Positive Rate which will naturally be lower than with a balanced dataset*
- Possible alternative metrics:
  - $Precision = \frac{True\ Positive\ (TP)}{TP + False\ Positive\ (FP)} = Predictive\ Positive\ Value$
  - $Recall = \frac{TP}{TP + False\ Negative\ (FN)} = Sensitivity$
  - $AUC\text{-}PR = Area\ Under\ Precision\text{-}Recall\ Curve$
  - $F1\text{-}Score = 2 * \frac{Precision * Recall}{Precision + Recall}$
- AUC examines the tradeoff between:
  - *Sensitivity*
  - $Specificity = \frac{True\ Negative\ (TN)}{TN + FP} = True\ Negative\ Rate = 1 - False\ Positive\ Rate$

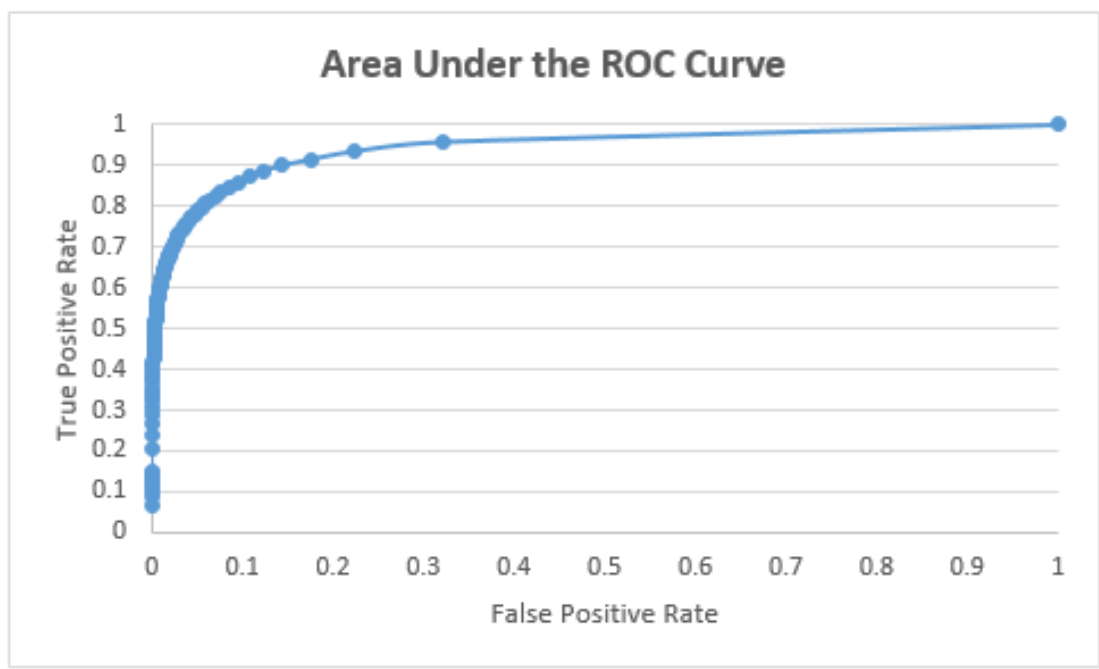
Sensitivity Analysis – A/55/45 Real Data

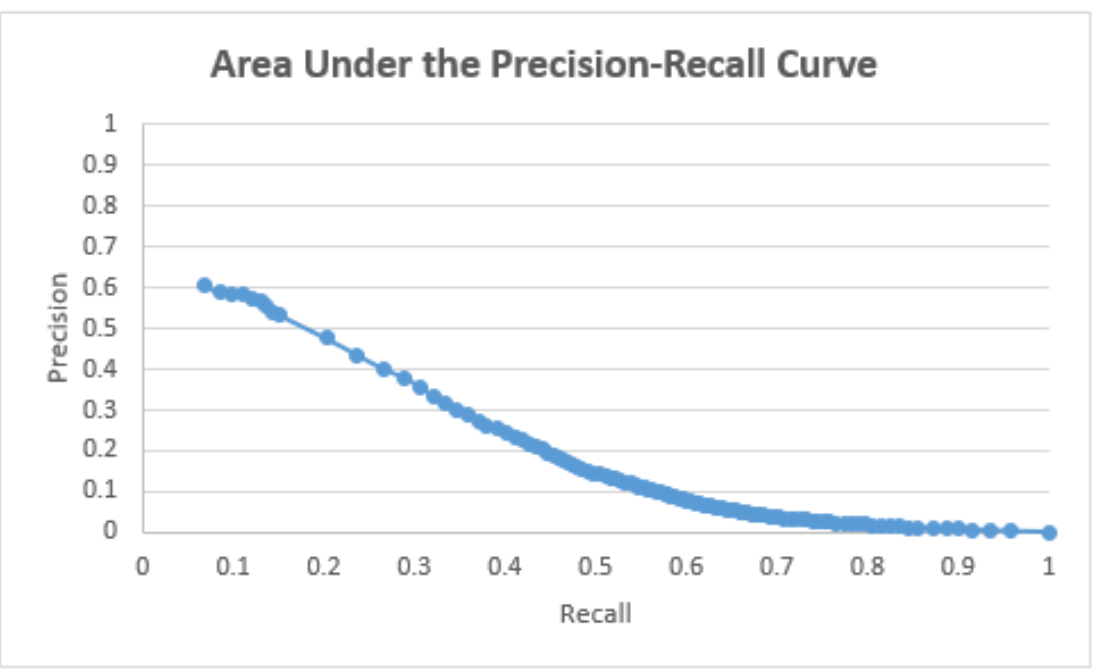
Gradient Boosted Tree Classifier – 2500 Boosting Rounds



Sensitivity Analysis – A/55/45 Real Data

Gradient Boosted Tree Classifier – 2500 Boosting Rounds





Short-Term Refinements

- Hyperparamater tuning of transfer learning classifier
- Soft Training
- Re-training fraud WGAN-GPs

Longer Term Extensions

- Data Augmentation – Blending real and synthetic data (single and multi-bank)
- Examination of other applications for non-fraud to fraud GAN transfer technique
- Weighted data transfer/training
- Extending to heterogeneous feature sets and/or missing data
- Testing on multiple datasets
- Using multiple GAN models and/or generated data samples as alternative to e.g. bagging in model development
- Bayesian interpretation of GANs as sampling from posterior predictive distribution
- Incorporate (ε, δ)-differential privacy

Bayesian Interpretation

If x represents our current sample and $\tilde{x}$ our new sample, then the posterior predictive distribution $\tilde{x}|x$ is written as:

$$p(\tilde{x}|x) = \int p(\tilde{x}|\theta)p(\theta|x)d\theta$$

where θ represents the distribution parameters and $p(\theta|x)$ the posterior distribution of θ with,

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

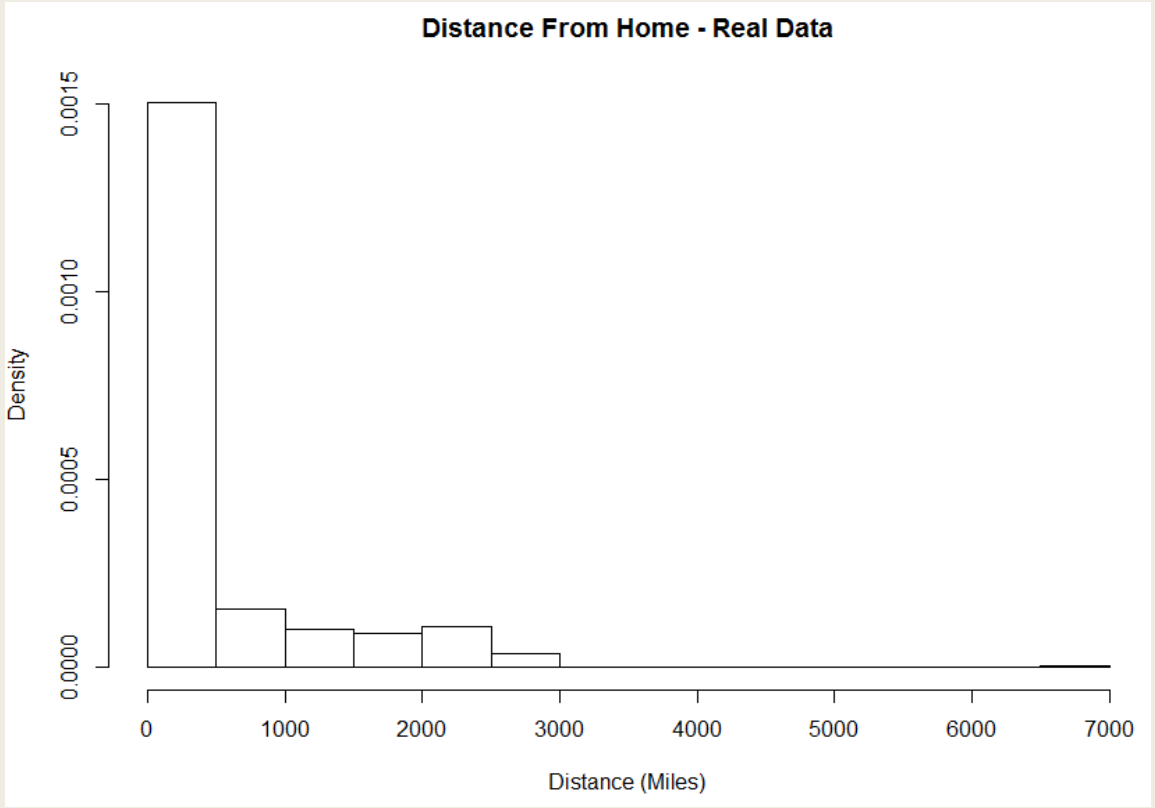
where $p(x|\theta)$ is the likelihood and $p(\theta)$ the prior distribution of θ

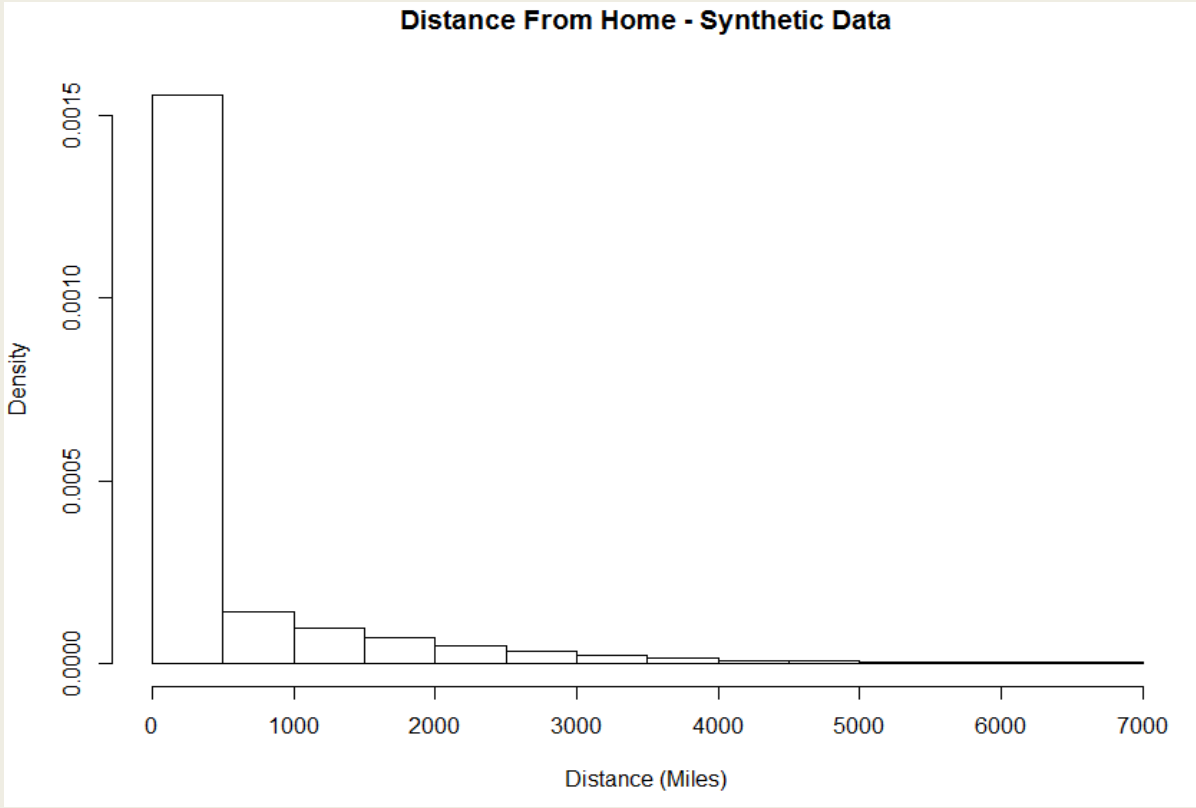
If we were to attach a prior distribution on the data itself, call it $p(z)$ then by training a GAN we are attempting to learn a function,

$$f: p(z) \rightarrow p(\tilde{x}|x)$$

Bayesian Interpretation - Example

Sampling From the Posterior Predictive Distribution





References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley D., Ozair, S., Courville, A. and Bengio, Y. "Generative Adversarial Nets." Proceedings of the International Conference on Neural Information Processing Systems (NIPS). 2014.

[2] Arjovsky, M., Chintala, S. and Bottou, L. "Wasserstein Generative Adversarial Networks." Proceedings of the 34th International Conference on Machine Learning. 2017.

[3] Gulrajani, I., Faruk, A., Arjovsky, M., Dumoulin, V. and Courville, A. "Improved Training of Wasserstain GANs." Proceedings of the International Conference on Neural Information Processing Systems (NIPS). 2017.

[4] Dwork, C. "Differential Privacy." Proceedings of the 33rd International Conference on Automata, Languages and Programming (ICALP). 2006.

[5] Yang, W., Adams, S., Beling, P., Greenspan, S., Rajagopalan, S., Velez-Rojas, M., Mankovski, S., Boker, S. and Brown, D. "Privacy Preserving Distributed Deep Learning and Its Application in Credit Card Fraud Detection." 17th IEEE International Conference on Trust, Security and Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science and Engineering (TrustCom/BigDataSE). 2018.

[6] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y. and Li, Y. "xgboost: Extreme Gradient Boosting". R package version 0.82.1. 2018. https://CRAN.R-project.org/package=xgboost.

References (Continued)

[7] Saito, T. and Rehmsmeier, M. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Data.” PLoS One. 2015.