Deep-Learning for Treatment Effect Estimation

Dingyi Lai, Farah Azham, Zhuocheng Xie

Applied Predictive Analytics (SoSe 2021)
Humbolt University of Berlin

June 20, 2021

Table of Contents

Table of Contents

Introduction

Treatment-Response Matrix with Four Customer Types

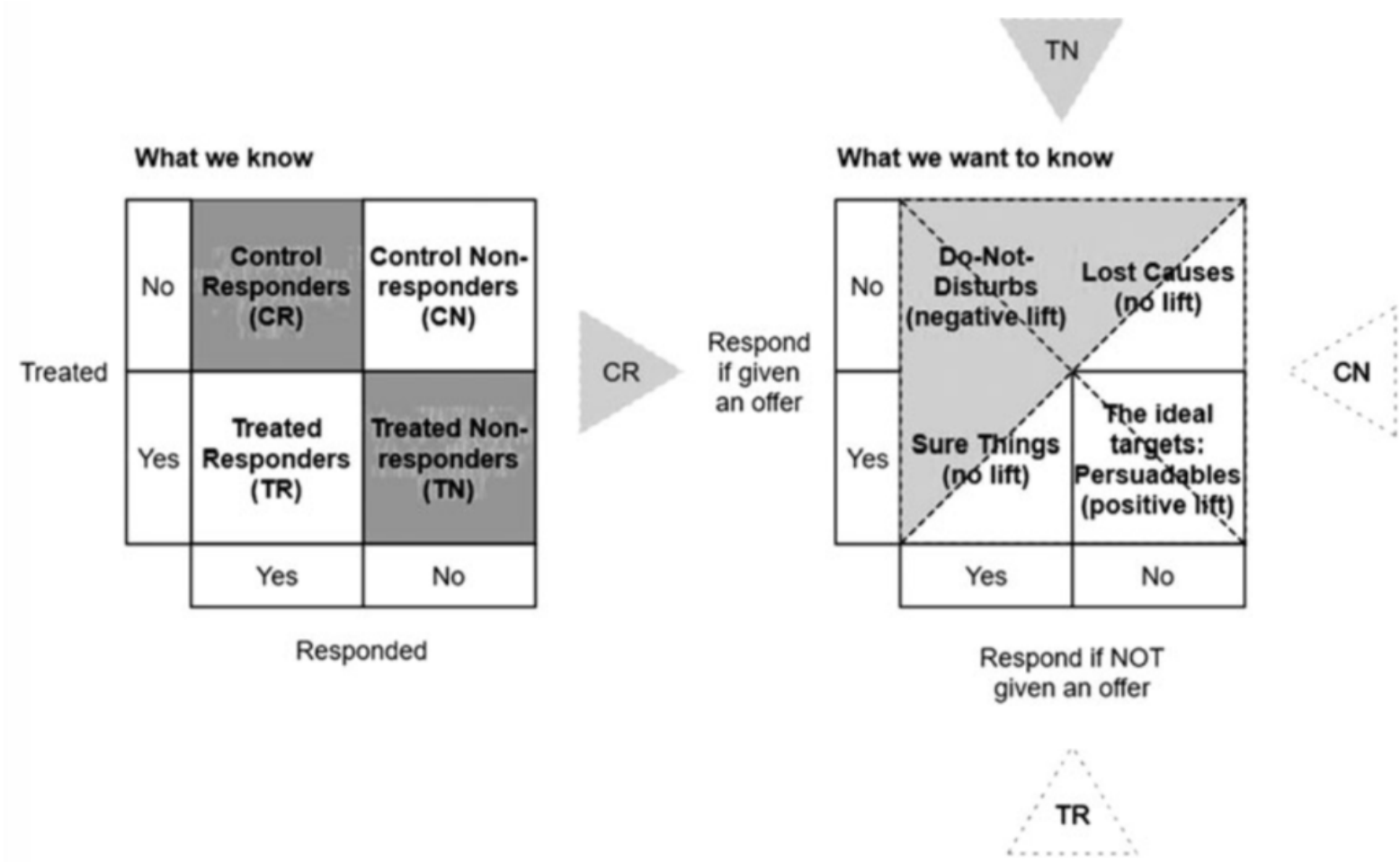

Figure 1.1: Conceptual Table (Devriendt, Moldovan and Verbeke, 2018)

Introduction

Types of Treatment Effects

Individual Treatment Effect (ITE)

$$\tau^{ITE} = \tau_i = Y_i(1) - Y_i(0)$$

Conditional Average Treatment Effect (CATE)

$$\tau^{CATE} = \tau(x) = E(Y_i(1) - Y_i(0)|X_i = x_i)$$

Average Treatment Effect (ATE)

$$\tau^{ATE} = E(\tau(x)) = E(Y_i(1) - Y_i(0))$$

Uplift Score

$$U(X_i) := p(Y_i|X_i, T_i = 1) - p(Y_i|X_i, T_i = 0)$$

Table of Contents

Data

We choose Kevin Hillstrom's dataset from E-Mail Analytics And Data Mining Challenge (Hillstrom 2008). The dataset consists of 64,000 records reflecting customers that last purchased within 12 months.

	recency	history_segment	history	mens	womens	zip_code	newbie	channel	segment	visit	conversion	spend
0	10	2) 100–200	142.44	1	0	Surburban	0	Phone	Womens E-Mail	0	0	0.0
1	6	3) 200–350	329.08	1	1	Rural	1	Web	No E-Mail	0	0	0.0
2	7	2) 100–200	180.65	0	1	Surburban	1	Web	Womens E-Mail	0	0	0.0
3	9	5) 500–750	675.83	1	0	Rural	1	Web	Mens E-Mail	0	0	0.0
4	2	1) 0–100	45.34	1	0	Urban	0	Web	Womens E-Mail	0	0	0.0

Concerning the indicator of treatment t, segment is appropriate. There are 3 levels:

- Mens E-Mail
- Womens E-Mail
- No E-Mail

Choosing Target variable

Among the 12 columns, there are 3 possible targets.

- Visit: 1/0 indicator,
  1 = customer visited website in the following two weeks
- Conversion: 1/0 indicator, 1 = customer purchased merchandise in the following two weeks
- Spend: actual dollars spent in the following two weeks

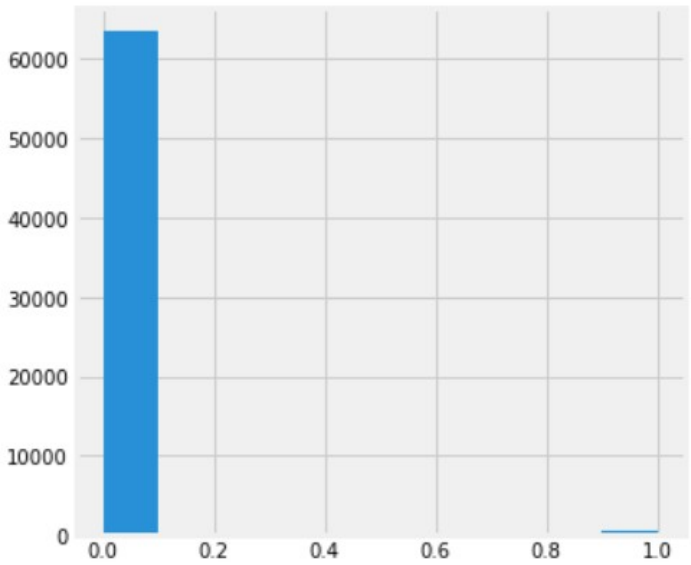

Figure 2.2: Plot of Conversion

Pie graph of customers

After determining the target variable, we make a pie plot in order to show the proportion of customers in **CR, TR, CN, TN** four groups.

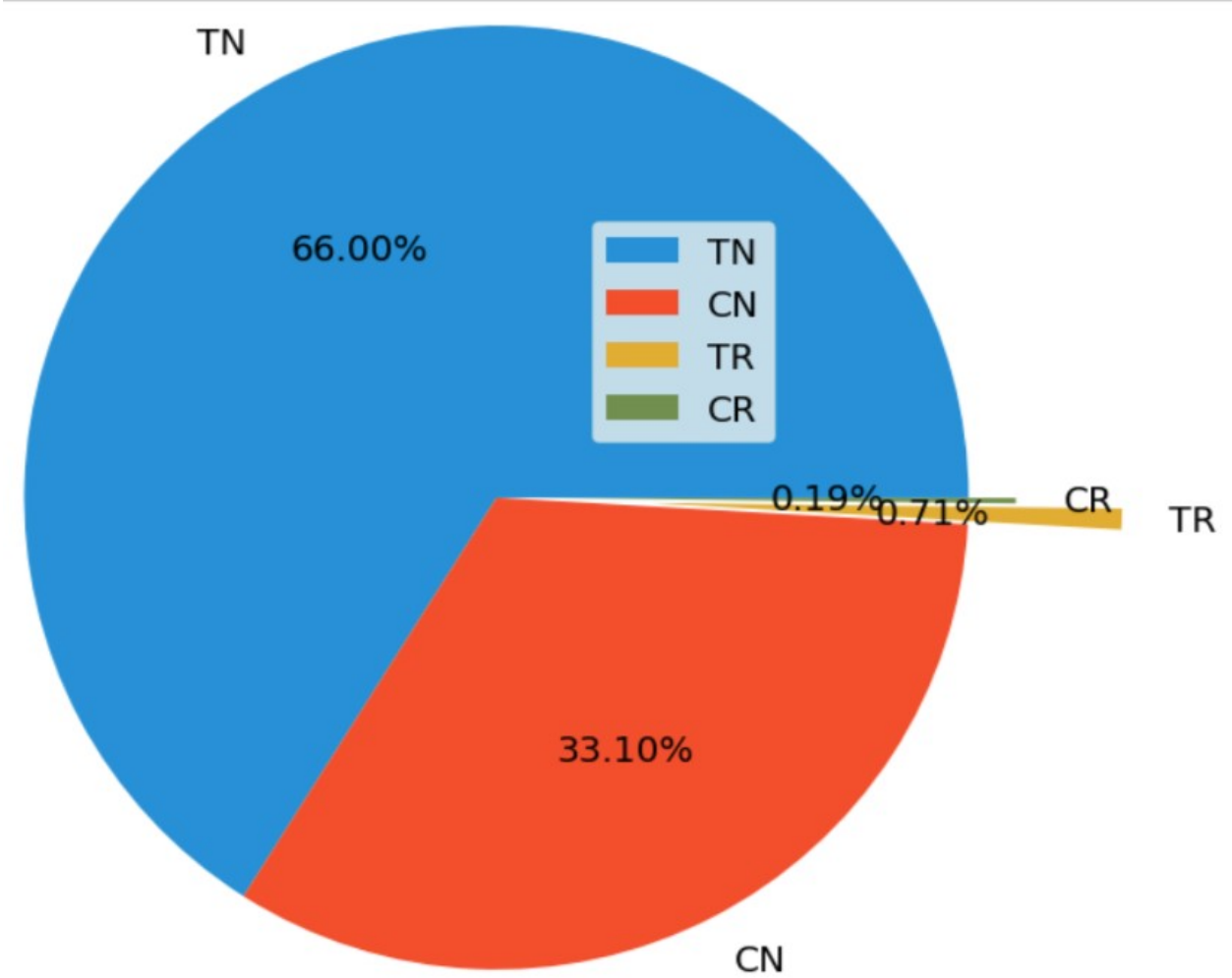

Figure 2.3: Pie Plot

Table of Contents

Different Methods for Estimation of Treatment Effects Considering Conversion

- Evaluation Metrics for Uplift Models
- Benchmark Model: S-learner and T-learner from Kunzel et al.(2019)
- Dragonnet from Shi, Blei and Veitch (2019)
- Causal Net from Farrell et al.(2018)

Evaluation Metrics for Uplift Models

- The **Area Under the Uplift Curve (AUUC)** calculates the area underneath the uplift curve compared with the area under the optimal curve.
- The **Qini coefficient** is a single number performance measure. It is the area between the uplift model's Qini curve and the random targeting line.

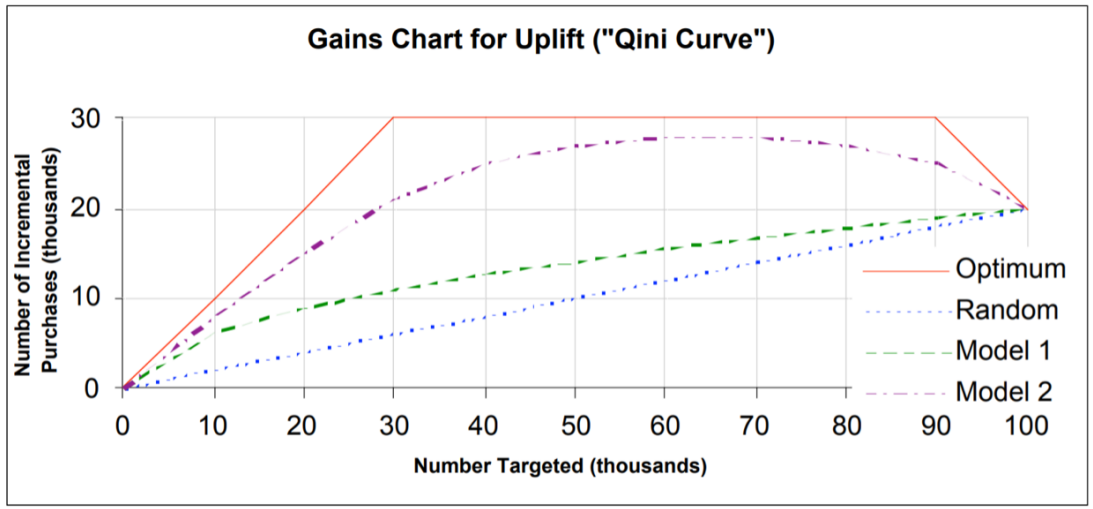

Figure 3.4: Gain Chart for Uplift-Curve (Radcliffe, 2007)

Benchmark Model

We choose S- and T-learner, two base-learners in meta-algorithms (Kunzel et al. 2019) as our benchmark models for estimating CATE.

T-learner

The T-learner is a tree-based method and includes two tree base learners in it, in order to estimate the response from treatment and control groups. There are two steps of obtaining T-learner:

- Use any supervised learning or regression to estimate

$$\mu_0(x) = E[Y(0)|X = x] \quad \mu_1(x) = E[Y(1)|X = x]$$

  With the treated observation, we can denote the estimator by $\hat{\mu_0}(x)$ and $\hat{\mu_1}(x)$.

- The CATE of T-learner is defined as:

$$\hat{\tau_T}(x) = \hat{\mu_0}(x) - \hat{\mu_1}(x)$$

Benchmark Model

S-learner

"S-learner", a "single estimator", estimates the outcome using all of the features and the treatment indicator, without giving the treatment indicator a special role.

The predicted CATE for an individual unit is then the difference between the predicted values when the treatment assignment indicator is changed from control to treatment, with all other features hold fixed. Thus, we estimate the combined response function,

$$\mu(x, w) := E[Y^{obs}|X = x,\ W = w],$$

using any base learner on the entire data set. Also, we denote the estimator as $\hat{\mu}$. The CATE of S-learner is defined as:

$$\hat{\tau}_s(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$$

Benchmark Model

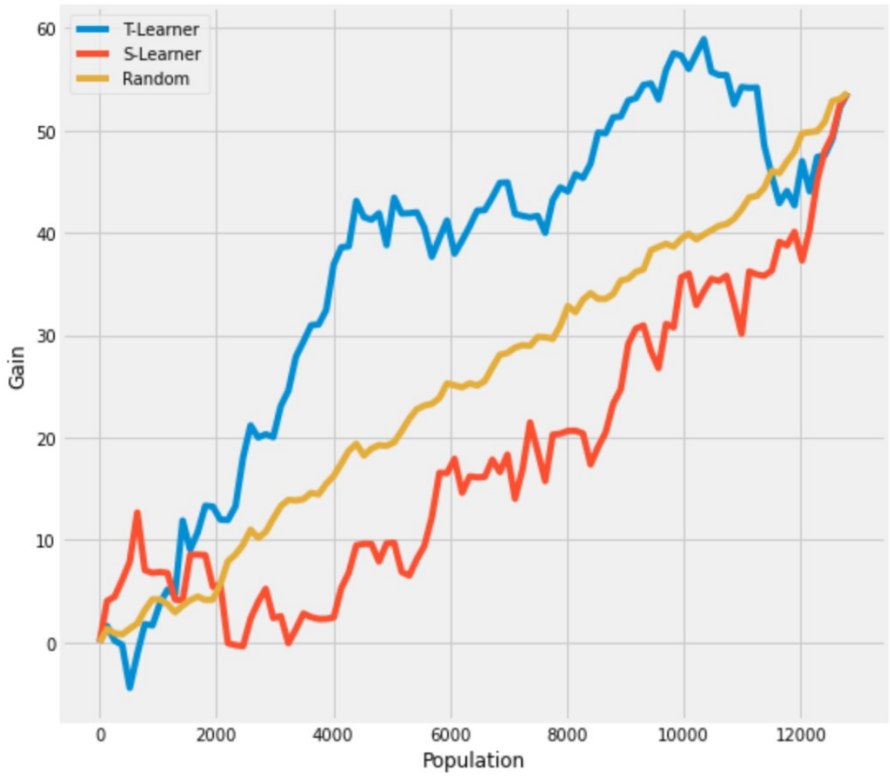

Figure 3.5: Qini Curves of Benchmark Models

Table 3.1: Results for Benchmark

Learner	AUUC	Qini
S-learner	0.005115	-0.090615
T-learner	0.677724	0.749913
random	0.472620	1.492023e-15

Dragonnet from Shi, Blei and Veitch (2019)

- Shi, Blei and Veitch (2019) introduce a three-headed neural network architecture that provides an end-to-end training to predict the conditional outcome and the propensity score.
- deep net to produce a representation layer $Z(X) \in \mathbb{R}^p$
- simple linear map (followed by a sigmoid) for the propensity score model $\hat{g}$
- Sufficiency of the Propensity Score

  $$\tau = E[E[Y|g(X), T = 1] - E[Y|g(X), T = 0]] \ ,$$

  where the propensity score is $g(x) = P(T = 1|X = x)$

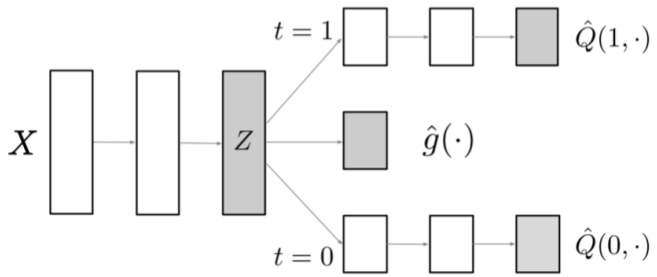

Figure 3.6: Dragonnet Architecture (Shi, Blei and Veitch, 2019)

Dragonnet from Shi, Blei and Veitch (2019)

- Estimated average treatment effect with conditional outcome

  $\hat{\tau}^Q = \frac{1}{n}\sum_i[\hat{Q}(1, x_i) - \hat{Q}(0, x_i)]$,

  where the conditional outcome is $Q(t, x) = E[Y|t, X = x], t \in 0, 1$.
- The CATE of Dragonnet is defined as:
  $\tau = E[E[Y|T = 1, X = x] - E[Y|T = 0, X = x]]$

- Objective Function:

  $\hat{\theta} = argmin_\theta \hat{R}(\theta; X)$, where

  $\hat{\mathrm{R}}(\theta; X) = \frac{1}{n}\sum_i[(Q^{nn}(t_i, x_i; \theta) - y_i)^2 + \alpha CrossEntropy(g^{nn}(x_i; \theta), t_i)]$
- The targeted regularization is a modification to the objective function used for neural network training. We train the model by minimizing the modified objective

  $\hat{\theta}, \hat{\epsilon} = argmin_{\theta,\epsilon}[\hat{R}(\theta; X) + \beta\frac{1}{n}\sum_i \gamma(y_i, t_i, x_i; \theta, \epsilon)]$

Dragonnet from Shi, Blei and Veitch (2019)

Empirical Results

Architecture that we choose:

- neurons per layer [190], ratio [0.7], learning rate [0.0006]
- targeted regularization [True]

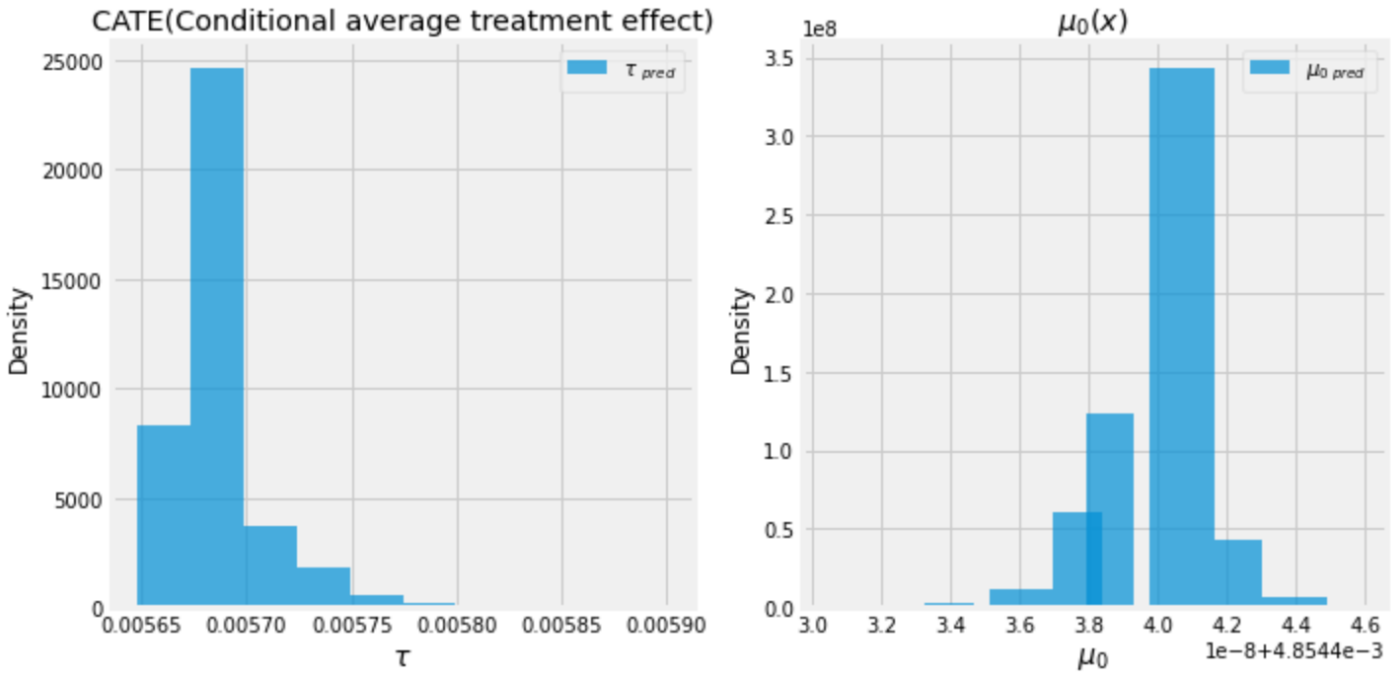

Figure 3.7: Distribution of CATE and μ_0

Dragonnet from Shi, Blei and Veitch (2019)

Evaluation

- AUUC Score is 0.6756 (compared to random 0.4726)
- Qini Score is 0.2019 (compared to random 1.492023e-15)

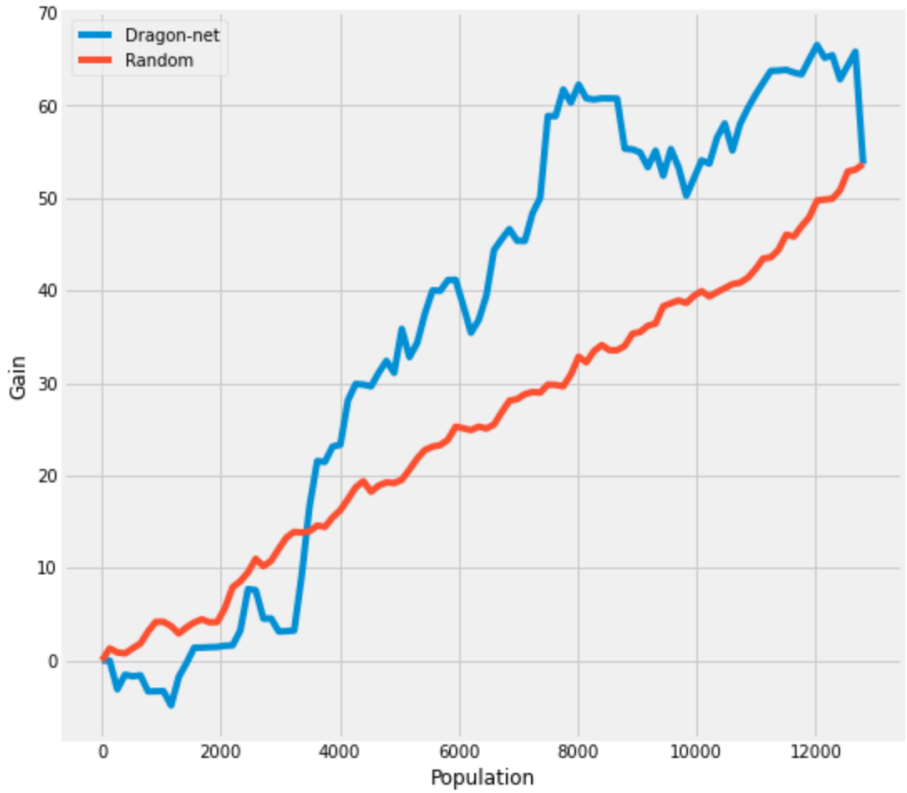

Figure 3.8: Qini Curve of Dragonnet

Causal Net from Farrell et al.(2018)

- Farrell et al.(2018) use ReLU as activation function $\sigma(x) = max(x, 0)$ to construct a feedforward neural network.

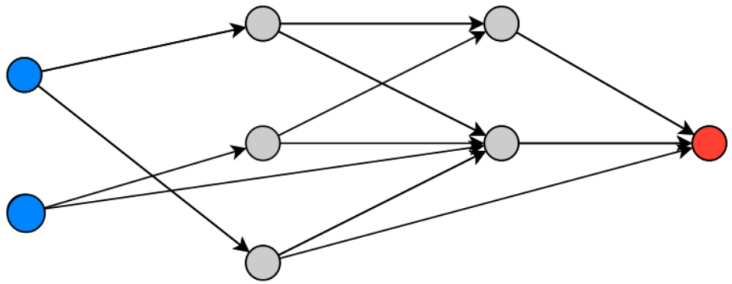

Figure 3.9: Causel Net architecture (Farrell et al., 2018)

- The estimates of the regression functions $\mu_t(x) = E[Y(t)|X = x]$, $t \in 0, 1$ is used for calculation of CATE: $\tau(x) = \mu_1(x) - \mu_0(x)$.
- Optimization function:

$$\begin{pmatrix} \hat{\mu_0}(x) \\ \hat{\tau}(x) = \hat{\mu_1}(x) - \hat{\mu_0}(x) \end{pmatrix} := \underset{\tilde{\mu_0}, \tilde{\tau}}{\text{argmin}} \sum_{i=1}^{n} \frac{1}{2}(y_i - \tilde{\mu_0}(x_i) - \tilde{\tau}(x_i)t_i)^2$$

Causal Net from Farrell et al.(2018)

Empirical Result

Architecture that we choose:

- hidden layer sizes is [60]; dropout rate is [0.5]
- optimizer is 'Adam'; learning rate is 0.0003

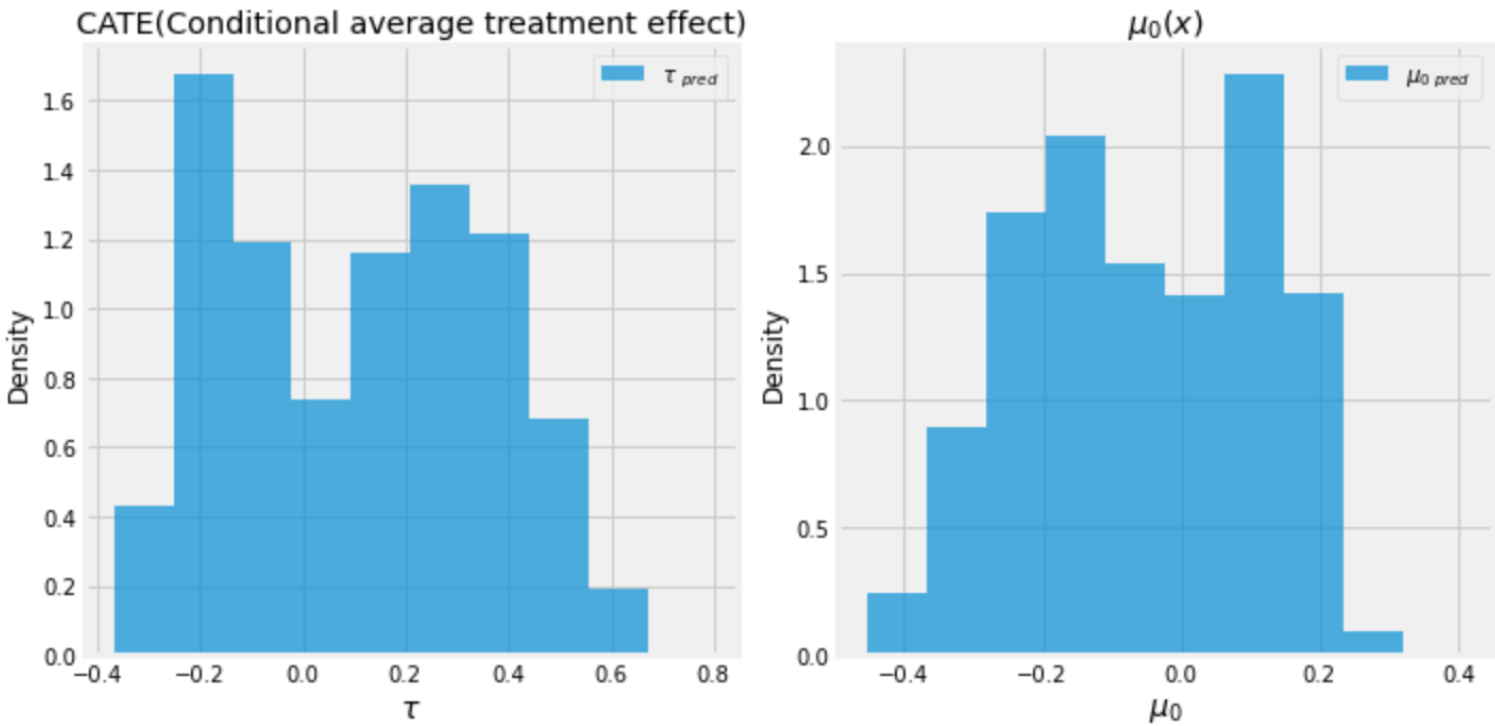

Figure 3.10: Distribution of CATE and μ_0

Causal Net from Farrell et al.(2018)

Evaluation

- AUUC Score is 0.7081 (compared to random 0.4726)
- Qini Score is 0.2361 (compared to random 1.4920e-15)

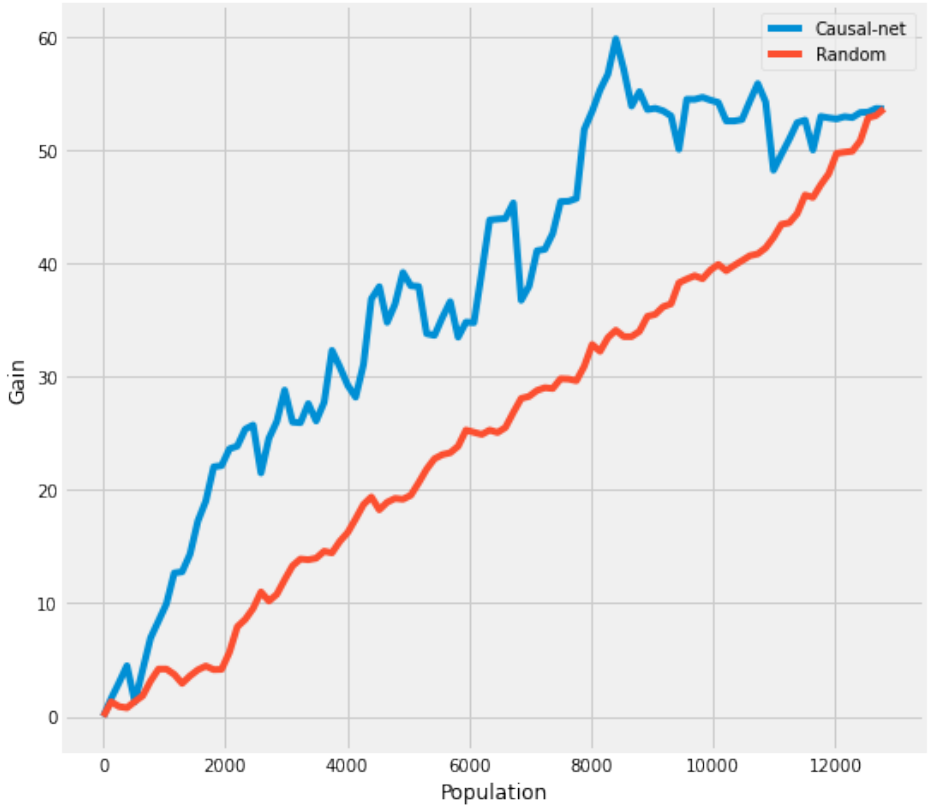

Figure 3.11: Qini Curve of Causal Net

Table of Contents

Tuning T-learner

- Since we do not know the suitable architecture for these algorithms and there is a risk to directly use the patterns provided by the paper, we decide to tune the model empirically.

Table 4.2: Tuning Results for T-learner

learning rate	ATE	AUUC	Qini
0.0002	0.000101456	0.677724	0.203848
0.0004	0.00020092	0.677724	0.203848
0.0006	0.000298598	0.677719	0.203841
0.0008	0.000394766	0.677719	0.203841
0.001	0.00048909	0.677719	0.203841

- T-learner is not sensitive to the parameter tuning, nonetheless, one could set learning rate to be 0.0004.

Tuning Dragonnet

- There are mainly three parameters that could be tuned: neurons per layer (npl), ratio and learning rate (lr).

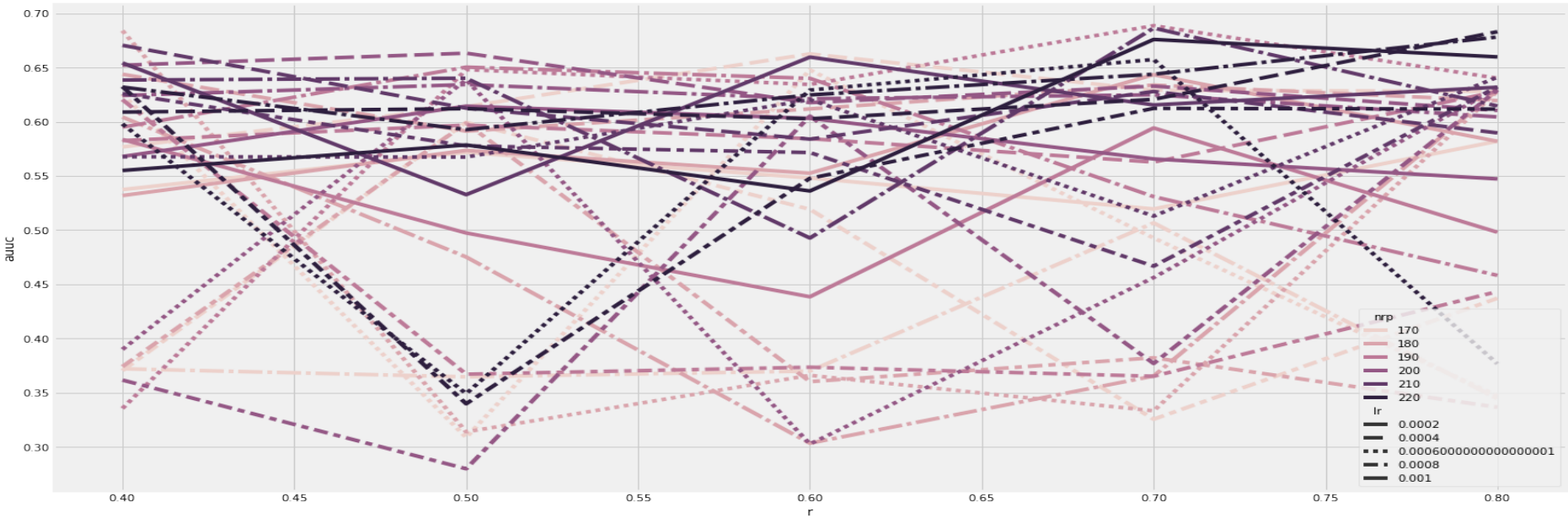

Figure 4.12: Tuning Results for Dragonnet

- Dragonnet is sensitive to the combination of different parameters, but there is no clear trend. The best pool among them is: npl=190, ratio = 0.7, lr=0.0006.

Tuning Causal Net

Tuning parameters in causal net is way more complicated, because there are several options, and especially for hidden layer size and dropout rate, we can even adjust the number of layers. First we choose the two-layers architecture, observing the trend of results from tuning. The architecture is hidden layer size = [30,i], dropout rate = [0.5,0].

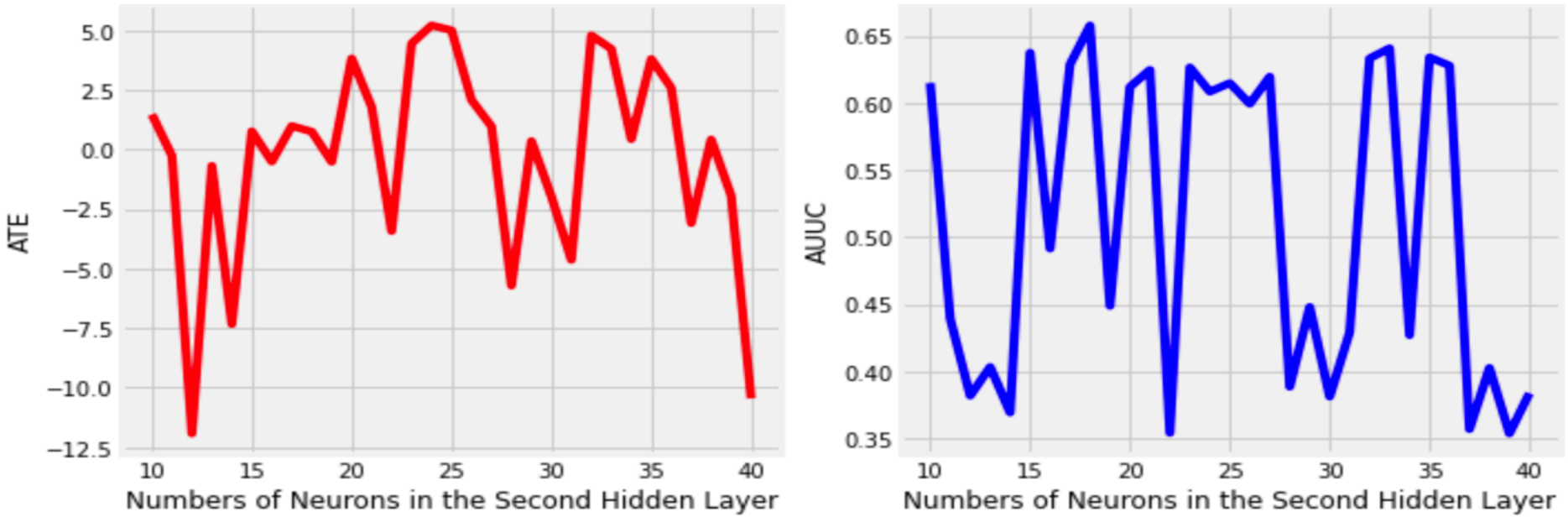

Figure 4.13: ATE and AUUC via Different Numbers of Neurons in the Second Hidden Layer i

Tuning Causal Net

Then we check the sensitivity of dropout rate:

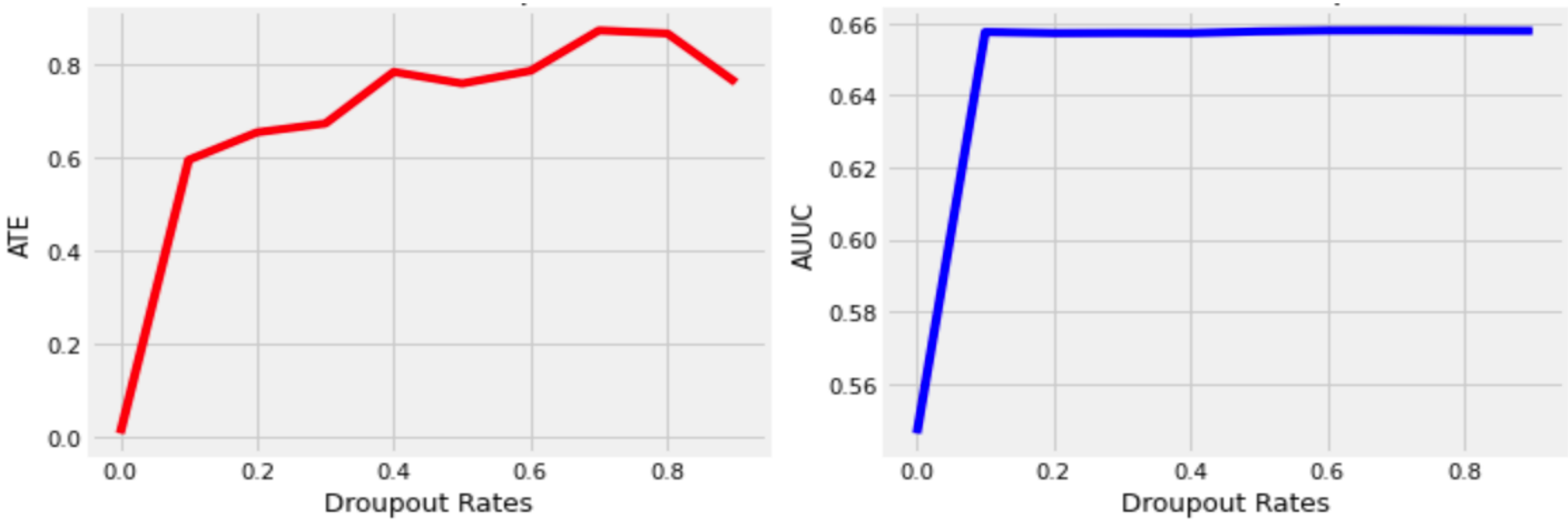

Figure 4.14: ATE and AUUC via Different Droupout Rates

ATE and AUUC are overall not too sensitive to dropout rate.

Tuning Causal Net

Next we try to delete one layer from that, keeping others the same. The result shows that one layer would be more possible to give a better result. The architecture for one-layer Causal Net is hidden layer size = [30], dropout rate = [0.5].

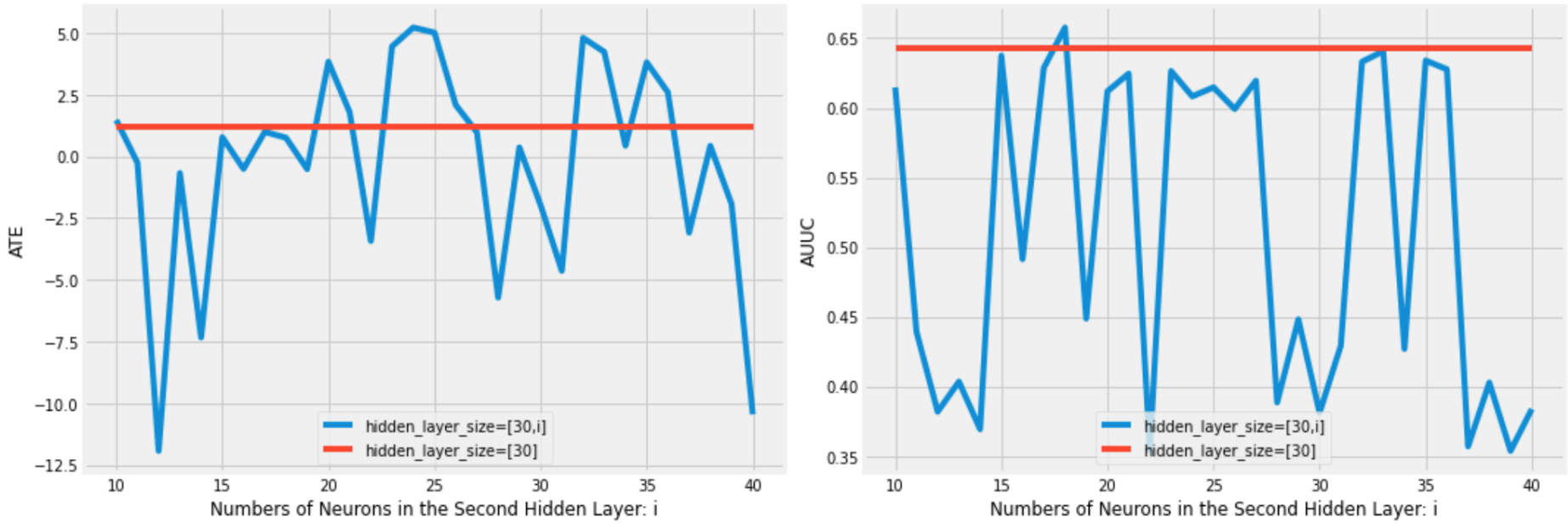

Figure 4.15: Comparison Between One-Layer and Two-Layers Architectures

Tuning Causal Net

Finally, we tune the one-layer architecture comprehensively empirically.

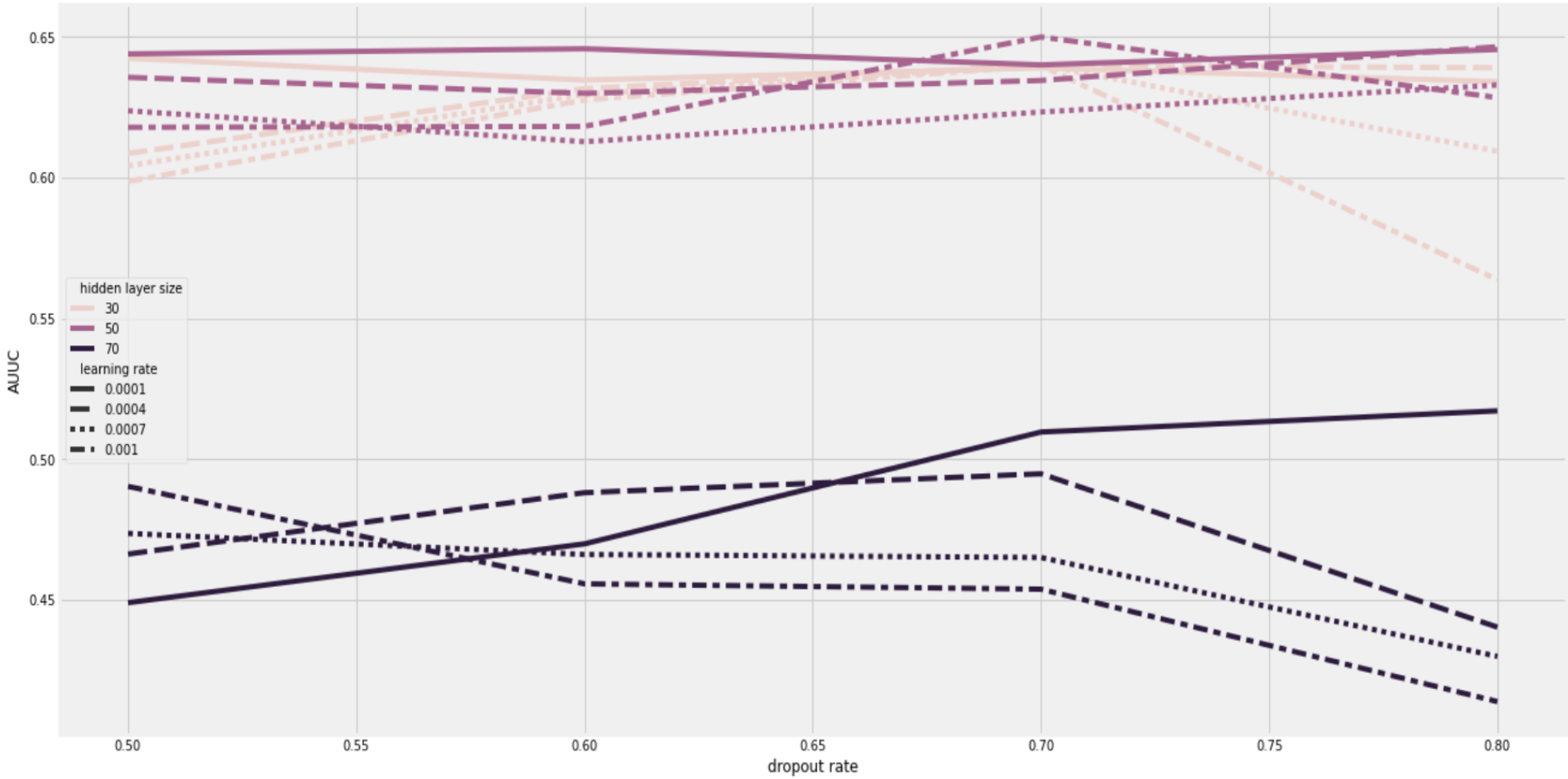

Figure 4.16: One-Layer Architecture Tuning

Conclusion from Tuning

Sensitivity Analysis

- AUUC in T-learner is insensitive to the learning rate;
- AUUC in Dragonnet is sensitive to the neurons per layer, ratio and learning rate, but it is difficult to see the trend; Nonetheless, we could draw a relatively better pool for analysis;
- Based on our data, it is more reasonable to use "spending" as target instead of "conversion", because ReLU activation function in causal net from Farrell et al.(2018) cannot constraint the output to be between 0 and 1. Also, ATE and AUUC in this method is sensitive to the hidden layer size and learning rate, but relatively not sensitive to dropout rate. Specifically, one-layer architecture might perform better than two-layers one;
- However, it needs more further research.

Table of Contents

Empirical Result

- The empirical ATE and AUUC from different methods are listed as below, using the relative reasonable parameter pools after tuning.

Table 5.3: Comparison between ATE, AUUC and Qini

Methods	ATE	AUUC	Qini
T-learner	0.000201	0.677724	0.203847
S-learner	0.005115	0.335156	-0.090615
Dragonnet	0.00569	0.675607	0.201902
Causal Net	0.109055	0.708083	0.236018
Random	-	0.472620	1.492023e-15

- Note that the result from Dragonnet is not stable, though we set the random seed before running the function. But generally, according to the AUUC, the performance is ranked as:

  Causal Net>T-learner≈Dragonnet>S-learner

Empirical Result

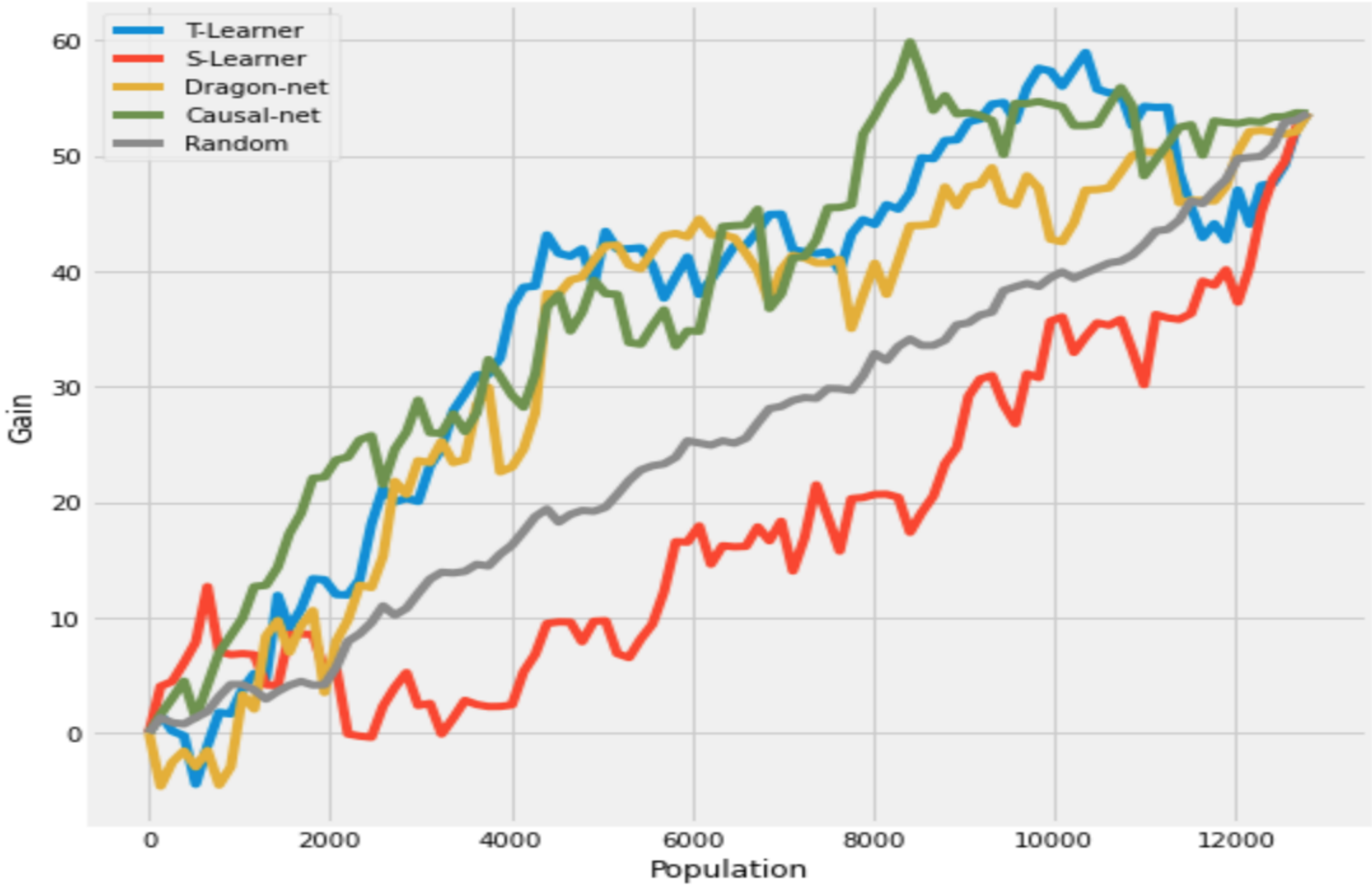

Figure 5.17: Comparison of Qini Curves among the Models

References

- Devriendt, F., Moldovan, D., and Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of- the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big Data, 6(1). 13-41.
- Shi, C., Blei, D. M. and Veitch, V. (2019). Adapting Neural Networks for the Estimation of Treatment Effects. ArXiv:1906.02120
- Farrell, M.H. and Liang, T. and Misra, S. (2018). Deep Neural Networks for Estimation and Inference. Econometrica. 89. 181–213.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10), 4156-4165.

References

- Kane, K., Lo, V. S., and Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. Journal of Marketing Analytics, 2(4). 218-238.
- Lai, L.Y.T. (2006). Influential marketing: A new direct marketing strategy addressing the existence of voluntary buyers. Master of Science thesis, Simon Fraser University School of Computing Science, Burnaby, BC, Canada.
- Radcliffe, N.J.(2007). Using control groups to target on predicted lift: Building and assessing uplift models, Direct Marketing Analytics Journal. 3. 14-21.

References

- Radcliffe, N.J. and Surry, P.D. (2011). Real-World Uplift Modelling with Significance-Based Uplift Trees. Portrait Technical Report, TR-2011-1.
- Rzepakowski, P. and Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems, 32(2), 303-327.