A Twin Neural Model for Uplift

Mouloud Belbahri, Olivier Gandouet, Alejandro Murua, Vahid Partovi Nia

June 17, 2021

Table of Contents

Table of contents

A retention campaign

Insurance companies are interested in retention strategies to minimize their attrition rate.

- Acquisition cost $\gg$ Retention cost
- Number of interactions with your insurer carrier is limited
  - Purchases
  - Life events
  - Claims
- Propensity model failure.
  - Higher risk of cancellation should be treated differently

A retention campaign

Insurance companies are interested in retention strategies to minimize their attrition rate.

- Acquisition cost $\gg$ Retention cost
- Number of interactions with your insurer carrier is limited
  - Purchases
  - Life events
  - Claims
- Propensity model failure.
  - Higher risk of cancellation should be treated differently

Table 1: Renewal rate by group for $n = 20997$ home insurance policies.

	Control	Called	Overall
Renewal rate	96.90%	96.50%	96.54%

Uplift modeling in a nutshell

- Marketing version of causal inference (Conditional Average Treatment Effect)

Uplift modeling in a nutshell

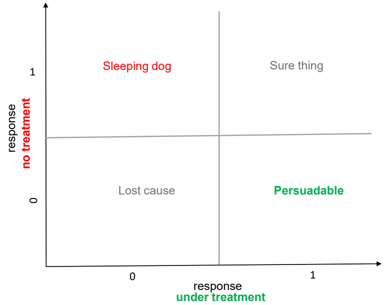
- Marketing version of causal inference (Conditional Average Treatment Effect)
- Given an action taken on a client uplift modeling aims to infer the impact of the action (or treatment) on a binary response.

Uplift modeling in a nutshell

- Marketing version of causal inference (Conditional Average Treatment Effect)
- Given an action taken on a client uplift modeling aims to infer the impact of the action (or treatment) on a binary response.
- Different types of clients.

Uplift modeling in a nutshell

- Marketing version of causal inference (Conditional Average Treatment Effect)
- Given an action taken on a client uplift modeling aims to infer the impact of the action (or treatment) on a binary response.
- Different types of clients.



Potential outcomes (Rubin)

Notation :

- X $= (X_1, \ldots, X_p)$ is the vector of pre-treatment characteristics;

Potential outcomes (Rubin)

Notation :

- X $= (X_1, \ldots, X_p)$ is the vector of pre-treatment characteristics;
- T is the binary treatment indicator variable (0 for control, 1 for treatment);

Potential outcomes (Rubin)

Notation :

- $\mathrm{X} = (X_1, \ldots, X_p)$ is the vector of pre-treatment characteristics;
- T is the binary treatment indicator variable (0 for control, 1 for treatment);
- $e(\mathbf{x}) = \Pr(T = 1 \mid \mathrm{X} = \mathbf{x})$ is the *propensity score*;

Potential outcomes (Rubin)

Notation :

- $\mathrm{X} = (X_1, \ldots, X_p)$ is the vector of pre-treatment characteristics;
- T is the binary treatment indicator variable (0 for control, 1 for treatment);
- $e(\mathbf{x}) = \Pr(T = 1 \mid \mathrm{X} = \mathbf{x})$ is the *propensity score*;
- Y_0 and Y_1 are the *potential outcomes* under control and treatment respectively.

Potential outcomes (Rubin)

Notation :

- $\mathrm{X} = (X_1, \ldots, X_p)$ is the vector of pre-treatment characteristics;
- T is the binary treatment indicator variable (0 for control, 1 for treatment);
- $e(\mathbf{x}) = \Pr(T = 1 \mid \mathrm{X} = \mathbf{x})$ is the *propensity score*;
- Y_0 and Y_1 are the *potential outcomes* under control and treatment respectively.

The uplift is defined as $u(\mathbf{x}) = \mathbb{E}(Y_1 - Y_0 \mid \mathrm{X} = \mathbf{x})$.

Potential outcomes (Rubin)

Notation :

- $\mathrm{X} = (X_1, \ldots, X_p)$ is the vector of pre-treatment characteristics;
- T is the binary treatment indicator variable (0 for control, 1 for treatment);
- $e(\mathbf{x}) = \Pr(T = 1 \mid \mathrm{X} = \mathbf{x})$ is the *propensity score*;
- Y_0 and Y_1 are the *potential outcomes* under control and treatment respectively.

The uplift is defined as $u(\mathbf{x}) = \mathbb{E}(Y_1 - Y_0 \mid \mathrm{X} = \mathbf{x})$.

Assumptions (Holland, 1986):

Assumption 1. (Overlap) For any $\mathbf{x}$, the true propensity score is strictly between 0 and 1, i.e., $0 < e(\mathbf{x}) < 1$.

Assumption 2. (Consistency) *Observed* outcome Y is represented using the potential outcomes and treatment assignment indicator as follows, $Y = TY_1 + (1 - T)Y_0$.

Unconfoundedness

Assumption 3. (Unconfoundedness) Potential outcomes (Y_0, Y_1) are independent ($\perp\!\!\!\perp$) of the treatment assignment indicator T conditioned on all pre-treatment characteristics, i.e., $(Y_0, Y_1) \perp\!\!\!\perp T|\mathrm{X}$.

Unconfoundedness

Assumption 3. (Unconfoundedness) Potential outcomes (Y_0, Y_1) are independent ($\perp\!\!\!\perp$) of the treatment assignment indicator T conditioned on all pre-treatment characteristics, i.e., $(Y_0, Y_1) \perp\!\!\!\perp T|\mathrm{X}$.

In the uplift framework (T binary randomized) we have the following equality:

$$\begin{aligned}
\mathbb{E}(Y_1 - Y_0 \mid \mathrm{X} = \mathbf{x}) &= \mathbb{E}(Y \mid T = 1, \mathrm{X} = \mathbf{x}) - \mathbb{E}(Y \mid T = 0, \mathrm{X} = \mathbf{x}) \\
&= \Pr(Y_i = 1 \mid \boldsymbol{X}_i = \mathbf{x}, T_i = 1) - \Pr(Y_i = 1 \mid \boldsymbol{X}_i = \mathbf{x}, T_i = 0) \\
&= m_{11}(\mathbf{x}) - m_{10}(\mathbf{x})
\end{aligned}$$

Table of contents

Existing methods

- The intuitive approach to model uplift is to build two classification models (Hansotia and Rukstales, 2001; Snowden et al., 2011; Austin, 2012) for $m_{1,1}$ and $m_{1,0}$

Existing methods

- The intuitive approach to model uplift is to build two classification models (Hansotia and Rukstales, 2001; Snowden et al., 2011; Austin, 2012) for $m_{1,1}$ and $m_{1,0}$
- Most active research in uplift modeling is in the direction of classification and regression trees (Breiman et al., 1984) where the majority are modified random forests (Breiman, 2001) where the uplift is estimated at the leaf node (Su et al., 2009; Chipman et al., 2010; Powers et al., 2018; Athey et al., 2019).

Posterior propensity scores

We define the posterior propensity scores as:

$$\Pr(T = 1 \mid Y = 1, \boldsymbol{X} = \mathbf{x}) = \frac{m_{11}(\mathbf{x})}{m_{11}(\mathbf{x}) + m_{10}(\mathbf{x})}, \tag{1}$$

$$\Pr(T = 1 \mid Y = 0, \boldsymbol{X} = \mathbf{x}) = \frac{m_{01}(\mathbf{x})}{m_{01}(\mathbf{x}) + m_{00}(\mathbf{x})}, \tag{2}$$

where $m_{yt}(\mathbf{x}) = \Pr(Y = y|\boldsymbol{X} = \mathbf{x}, T = t)$

These probabilities are connected to the relative risk and are "observable":

$$\mathrm{RR}(\mathbf{x}) = \frac{\Pr(Y = 1 \mid \boldsymbol{X} = \mathbf{x}, T = 1)}{\Pr(Y = 1 \mid \boldsymbol{X} = \mathbf{x}, T = 0)} = \frac{m_{11}(\mathbf{x})}{m_{10}(\mathbf{x})}. \tag{3}$$

The uplift loss function

Let $p_{yt} \stackrel{\text{def}}{=} p_{yt}(\mathbf{x}) = m_{yt}/(m_{y1} + m_{y0})$. We define the uplift loss function as follows:

$$\ell(\mathbf{y}, \mathbf{t} \mid \mathbf{x}) = -\frac{1}{n} \sum_{i=1}^{n} \Big(y_i \log m_{1t_i} + (1 - y_i) \log m_{0t_i} + t_i \log p_{y_i 1} + (1 - t_i) \log p_{y_i 0} \Big)$$

The uplift loss function

Let $p_{yt} \stackrel{\text{def}}{=} p_{yt}(\mathbf{x}) = m_{yt}/(m_{y1} + m_{y0})$. We define the uplift loss function as follows:

$$\ell(\mathbf{y}, \mathbf{t} \mid \mathbf{x}) = -\frac{1}{n} \sum_{i=1}^{n} \Big(\underbrace{y_i \log m_{1t_i} + (1 - y_i) \log m_{0t_i}}_{\text{conditional mean}} + \underbrace{t_i \log p_{y_i 1} + (1 - t_i) \log p_{y_i 0}}_{\text{posterior propensity}} \Big)$$

A twin neural model for uplift

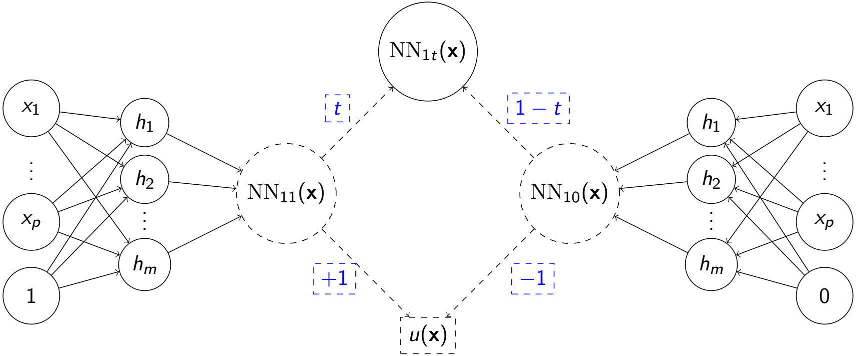


Figure 1: A twin neural model for uplift. The inputs contain the covariates vector $\mathbf{x}$ and, for the left sub-component, the treatment variable fixed to 1. The treatment variable is fixed to 0 for the right sub-component. The sub-components output the predicted conditional means for treated ($NN_{11}(\mathbf{x})$) and for control ($NN_{10}(\mathbf{x})$).

Table of contents

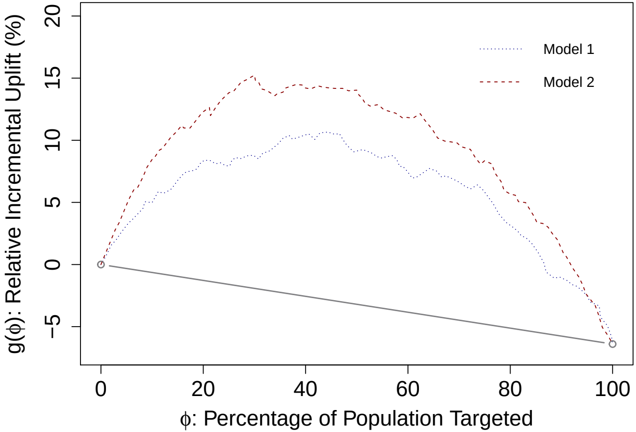
Evaluation metric: Qini

For a given model, let $\hat{u}_{(1)} \geq \hat{u}_{(2)} \geq ... \geq \hat{u}_{(n)}$ be the sorted predicted uplifts. Let $\phi \in [0, 1]$ be a given proportion and let $N_\phi = \{i : \hat{u}_i \geq \hat{u}_{\lceil \phi n \rceil}\} \subset \{1, \ldots, n\}$ be the subset of observations with the $\phi n \times 100\%$ highest predicted uplifts $\hat{u}_i$ (here $\lceil s \rceil$ denotes the smallest integer larger or equal to $s \in \mathbb{R}$). The *Qini curve* is defined as a function f of the fraction of population targeted ϕ, where

$$f(\phi) = \frac{1}{n_t} \left(\sum_{i \in N_\phi} y_i t_i - \sum_{i \in N_\phi} y_i (1 - t_i) \left\{ \sum_{i \in N_\phi} t_i / \sum_{i \in N_\phi} (1 - t_i) \right\} \right),$$

where $n_t = \sum_{i=1}^n t_i$ is the number of treated customers, with $f(0) = 0$ and $f(1)$ is the average treatment effect (ATE)

Evaluation metric: Qini



Simulations

Inspired by the simulations of Powers et al. (2018).

Our implementation	Scenarios			
Model	1	2	3	4
2-hidden layers *Twin*$_{\mathrm{NN}}$	**1.59**	2.36	1.38	**3.39**
Open-source implementation				
Qini-based (Belbahri et al., 2021)	1.02	2.68	1.09	2.94
Causal Forest (Athey et al., 2019)	0.75	2.22	0.94	2.79
Causal Forest (Honest) (Athey et al., 2019)	0.75	2.51	1.14	3.07
Uplift Random Forest (KL) (Guelman et al., 2012)	0.74	2.52	1.01	2.19
Uplift Random Forest (ED) (Guelman et al., 2012)	0.68	2.42	0.99	2.33
R-Learner (`XGboost`) (Nie and Wager, 2020)	0.76	2.63	**1.40**	2.12
R-Learner (`lasso`) (Nie and Wager, 2020)	0.66	2.75	0.87	2.83
X-Learner (`XGboost`) (Künzel et al., 2019)	0.72	2.57	1.31	2.37
X-Learner (`lasso`) (Künzel et al., 2019)	0.77	**2.78**	0.77	2.91

Table 2: Summary: models comparison in terms of $\hat{q}_{\mathrm{adj}}$ averaged on the test set over 20 runs. Note that the maximum standard-error is 0.15; we do not report them to simplify the Table.

Table of contents

Discussion

- Generalization to observational studies
- Architecture selection
- Theoretical development

Bibliography I

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based g-computation. *Multivariate behavioral research*, 47(1):115–135.

Belbahri, M., Murua, A., Gandouet, O., and Nia, V. P. (2021). Qini-based Uplift Regression. *Annals of Applied Statistics*.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Bibliography II

Guelman, L., Guillén, M., and Pérez-Marín, A. M. (2012). Random Forests for Uplift Modeling: an Insurance Customer Retention Case. In *Modeling and Simulation in Engineering, Economics and Management*, pages 123–133. Springer.

Hansotia, B. and Rukstales, B. (2001). Direct marketing for multichannel retailers: Issues, challenges and solutions. *Journal of Database Marketing and Customer Strategy Management*, 9(3):259–266.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

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.

Nie, X. and Wager, S. (2020). Quasi-oracle Estimation of Heterogeneous Treatment Effects. *Biometrika*.

Bibliography III

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*, 173(7):731–738.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158.