A Large Scale Benchmark for Uplift Modeling

Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini

A Large Scale Benchmark for Uplift Modeling

Eustache Diemert*
CAIL
e.diemert@criteo.com

Artem Betlei
CAIL & Université Grenoble Alpes
a.betlei@criteo.com

Christophe Renaudin
CAIL
c.renaudin@criteo.com

Massih-Reza Amini
Université Grenoble Alpes
massih-reza.amini@imag.fr

ABSTRACT

Uplift modeling is an important yet novel area of research in machine learning which aims to explain and to estimate the causal impact of a treatment at the individual level. In the digital advertising industry, the treatment is exposure to different ads and uplift modeling is used to direct marketing efforts towards users for whom it is the most efficient [1]. To foster research in this topic we release a publicly available collection of 25 million samples from a randomized control trial, scaling up previously available datasets by a healthy 590x factor. We provide details on the data collection and sanity checks performed that allow the use of this data for counter-factual prediction. We formalize the task of uplift prediction that could be performed with this data, along with the relevant evaluation metrics. Finally we show that the dataset size makes it now possible to reach statistical significance when evaluating baseline methods on the most challenging target.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Computing methodologies** → **Machine learning**; **Supervised learning**;

KEYWORDS

Uplift Prediction, Causal Inference, Digital Advertising, Individual Treatment Effect

ACM Reference Format:
Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. 2018. A Large Scale Benchmark for Uplift Modeling. In *Proceedings of AdKDD & TargetAd (ADKDD'18).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Performance advertising has become a very successful model of programmatic advertising where advertisers payment is based on delivered value as measured by events of interest (mostly site visits and conversions). The industry standard practice is to attribute conversions to advertising events such as ad displays and clicks. However, such *after the fact* attribution cannot be made *causal* unless the analyst assumes that all variables potentially influencing the outcome are observed, referred to as unconfoundedness in the literature [2], a very strong assumption in nowadays complex advertising landscape [3]. On the other side advertisers perform *incrementality tests*, a particular randomized control trial where part of the users are prevented from being targeted to measure the causal effect of advertising campaigns [4]. This approach is immune to unobserved confounders but is limited to the estimation of the Average Treatment Effect (ATE). The corresponding task at the individual level is uplift modeling also called Individual Treatment Effect (ITE) estimation especially in the observational causality literature.

Previous works on uplift/ITE prediction were evaluated on synthetic data [5], non-counter-factual data [6], small, medical data [7] [8] or closed data [9], all of which incur particular limitations in interpretability or reproducibility. For instance, non-causal data introduces a bias as the treatment is no more independent from the covariates, hence limiting the interpretability unless special assumptions and methods from observational causality are introduced into the picture. In 2008 a medium-scale, counter-factual dataset has been released [10] and used in publications [11] [12] [13] [14]. As we will discuss in Section 3 and 5 the CRITEO-UPLIFT1 dataset can be seen as a continuation of this pioneering work that opened the door to more realistic, reproducible research. In particular we scale it up in terms of data size but also challenge as the imbalance in treatment/control is increased and average response level is much lower.

2 PROBLEM FORMULATION

The causal uplift $U(x)$ is the expected difference in outcome *should* the individual be selected to take the treatment or not. We formalize it using Pearl's causal inference framework [15] in Equation 1.

$$U(x) = \mathbb{E}[Y|X = x, do(T = 1)] - \mathbb{E}[Y|X = x, do(T = 0)] \quad (1)$$

Conversely, the conditional uplift $u(x)$ in Equation 2 is the expected difference in outcome *when* the individual has taken the treatment or not: that is when we observe it after the fact.

$$u(x) = \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0] \quad (2)$$

Causal and conditional uplifts are equivalent if treatment was administered at random:

$$T \perp\!\!\!\perp X \Rightarrow U(x) \equiv u(x)$$

*corresponding author

ADKDD'18, August 2018, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Note that it is always possible to learn a predictor of $u(x)$ using traditional approaches in supervised learning, even though we only observe treatment and its outcome from a given experiment. However, in order to interpret the uplift predictions as causal (especially for taking actions like exposing users to ads or taking medicine) the model must be learned on data for which $U(x) \equiv u(x)$. In particular, data collected in a counter-factual manner, for example in a randomized control trial (A/B test) fits that purpose. Therefore we assume a dataset composed of i.i.d. samples of the joint covariates X, label Y and treatment T variables:

$$\mathcal{D} = \{X_i, Y_i, T_i\}_{i=1\dots n} \; ; \; T_i \perp\!\!\!\perp X_i, \forall i$$

Learning algorithms have access to $\mathcal{D}$ and can learn any distributions (we will see that there are multiple possible choices). We consider a binary outcome: at inference time the model performs a prediction of the form:

$$\hat{u}(x) = \hat{P}_T(Y = 1|X = x) - \hat{P}_C(Y = 1|X = x)$$

that is with the same x the model predicts the difference between two potential outcomes $P_T(Y|X = x)$ and $P_C(Y|X = x)$, if the subject is treated or not, respectively.

3 DATASET

The CRITEO-UPLIFT1 dataset is constructed by assembling data resulting from several incrementality tests, a particular randomized trial procedure where a random part of the population is prevented from being targeted by advertising. The dataset consists of 25M rows, each one representing a user with 12 features, a treatment indicator and 2 binary labels (visits and conversions). Positive labels mean the user visited/converted on the advertiser website during the test period (2 weeks). The global treatment ratio is 84.6%. It is usual that advertisers keep only a small control population as it costs them in potential revenue. For privacy reasons the data has been sub-sampled non-uniformly so that the original incrementality level cannot be deduced from the dataset while preserving a realistic, challenging benchmark. Feature names have been anonymized and their values randomly projected so as to keep predictive power while making it practically impossible to recover the original features or user context. The dataset is available publicly from the Criteo datasets Web page[1].

3.1 Pitfalls & Sanity Checks

Collecting such a dataset can be challenging as there are potentially many pitfalls that can impair the causal interpretation of the data. A particular characteristic of the current generation of production systems is that they target users dynamically based on observed interactions over time [16] . That means that even in a randomized control trial (A/B test) setup interactions with the system influence subsequent ad exposure via adjustments of the bids based on user reactions. In particular, interactions after the first one are influenced both by the treatment and by the first interactions. This situation calls for either considering only the first interaction of a user during an A/B test or to log the user variables at the start of the test and observe the reward during the test. We have chosen the latter solution as it enforces logging of features at the same time for all users, minimizing chances to observe temporal shifts in their distribution for different reasons like sales periods or production platform evolutions.

We performed a first sanity check: no users in the control population should be exposed to ads during the test. We observed a feeble 0.001% ratio of non-compliance to this assumption and chose to remove these users from the dataset.

A second sanity check is that the treatment should be independent of the features: $T \perp\!\!\!\perp X$. A convenient way to verify this assumption is to perform a Classifier 2 Sample Test (C2ST) [17]: a classifier trained to predict treatment should not do better than chance level. The distribution of H_0 in this case is obtained by computing the test loss of classifiers trained to predict random splits in the data. Figure 1 illustrates the null distribution of the Hamming loss.

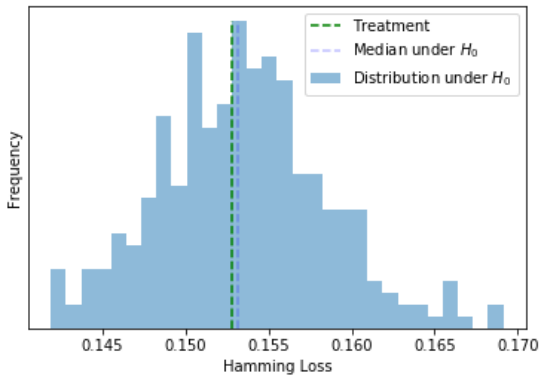

Figure 1: CRITEO-UPLIFT1**: Empirical distribution of Hamming loss under H_0 (blue), median under H_0 (blue, dotted) and treatment classifier loss (green, dotted). Predicting treatment is as good as predicting a random split.**

Table 1 gives the result of the test. Note that chance level is not 0.5 but rather $1 - \bar{T} \approx .15$, that is the performance of a dummy classifier always predicting $\bar{T}$. The empirical loss of the learned treatment classifier is very close to the dummy one from H_0, which is reflected by a high p-value for the one-sided test.

Median Random Loss	Treatment Loss	p-value
0.15312	0.15284	0.47176

Table 1: CRITEO-UPLIFT1**: Result of C2ST on treatment predictability with 300 resamples using Hamming loss. The p-value doesn't allow to reject H_0 and validates that $T \perp\!\!\!\perp X$**

A third level of sanity check is to make sure that logged features are informative and relevant for predicting outcomes (visit and conversion). This is not necessarily trivial as we sampled features that were technically easy to log and anonymized them. Table 2 presents the performance (as measured by log-loss) of classifiers learned on

[1] http://cail.criteo.com/criteo-uplift-prediction-dataset/

the outcomes for treatment, control and the whole dataset. The non-trivial improvement over a dummy baseline indicates that features are indeed informative for the task. We also measure positive, substantial mutual information of each individual feature and outcome but don't provide detailed results to reduce clutter.

	log-loss improvement (%)
conversion	59.80
conversion (control)	43.00
conversion (treated)	49.52
visit	51.70
visit (control)	55.02
visit (treated)	53.41

Table 2: CRITEO-UPLIFT1**: Improvement over the log-loss of a dummy classifier for different outcomes and groups. Baseline is a classifier predicting the average outcome, and improvement is relative to baseline. Big improvements indicate that the features have predictive power.**

3.2 Comparison to Existing Datasets

We now compare the CRITEO-UPLIFT1 dataset to the second largest and most popular uplift prediction dataset: HILLSTROM [10]. We don't consider datasets from health and medecine as their size is unrealistically small for our application (in the range of 100's). HILLSTROM contains results of an e-mail campaign for an Internet based retailer. The dataset contains information about 64,000 customers who last purchased within at most twelve months. The customers were involved in an e-mail test and were uniformly assigned to receive an e-mail campaign featuring men's merchandise / women's merchandise or not receive an e-mail. We report results on the Women's merchandise e-mail versus no e-mail split as in previous research [12] (the other partition is similar).

Table 3 summarizes the two datasets. Criteo dataset is much larger (a 590x factor). The treatment is not balanced, a typical situation when running incrementality tests on live campaigns. Of course one could down-sample the control population of Hillstrom to simulate such a situation but the data size would be ridiculously small. Both datasets pass the treatment independence test: barely for Hillstrom and with a high margin for Criteo. Although the response levels are 3x less in Criteo for both targets (making it more challenging) provided features are informative enough to learn a good predictor. We don't recommend to use the conversion target for Hillstrom as there are only 311 positive examples. The uplift levels are comparable.

4 EVALUATION

The dataset was collected and prepared with uplift prediction in mind as the main task. Additionally we can foresee related usages such as modeling, answering questions such as for instance: Is the response uniformly affected by the treatment or are there rich interactions between features and treatment. Similarly, and in the spirit of [18] one could also explore heterogeneity of treatment. Finally this dataset could also be used as benchmark for observational causality methods [3] as it contains the randomized causal effect plus covariates and exposure to compute propensity scores. In the rest of this section we focus on the uplift prediction task.

Metric	HILLSTROM	CRITEO-UPLIFT1
Size	42,693	25,309,483
Treatment Ratio	.50	.85
Average Visit	.12883	.04132
Average Conversion	.00728	.00229
Relative Avg. Uplift (visit, %)	42.6	68.7
Relative Avg. Uplift (conversion, %)	54.3	37.2
Treatment Independence	.056	.470
Learnability (visit, %)	45.81	51.70
Learnability (conversion, %)	48.85	59.80

Table 3: Summary of datasets characteristics. Independence is measured as C2ST p-value. Learnability is estimated as log-loss improvement over baseline

4.1 Metrics

The fundamental problem of evaluation for uplift prediction is that we observe only one of the potential outcomes $y|do(t=1)$ or $y|do(t=0)$, preventing the use of a direct loss on the uplift. Therefore one is bound to estimate if the predicted uplift is reasonable for paired groups of samples. For instance, consider the difference between two potential uplifts:

$$\begin{aligned} U(x_j) - U(x_i) = y_j|do(t=1) - y_i|do(t=1) \\ - \big(y_j|do(t=0) - y_i|do(t=0)\big) \end{aligned} \tag{3}$$

Even though we can only observe two variables out of four we know the sign of the difference in some cases; e.g. $y_j|do(t=1)=1$ and $y_i|do(t=1)=0$, then the difference is positive:

$$\begin{aligned} U(x_j) - U(x_i) = 1 - 0 \\ - \big(y_j|do(t=0) - y_i|do(t=0)\big) \geq 0 \end{aligned} \tag{4}$$

Two main metrics have been proposed[2] and both rely on the mentioned strategy. They also draw on ROC curves traditionally used for classifier evaluation and can be seen as natural extensions in the uplift case. The general idea is to rank individuals in the evaluation set according to their *predicted* uplift and cumulatively sum over them a measure of the actual, group-wise uplift. The intuition is that a good model should be able to select individuals with positive outcome in the treated group and negative outcome in the control group first.

Notations. For a given model let π be the ordering of the dataset satisfying $\hat{u}^{\pi}(x_i) \geq \hat{u}^{\pi}(x_j),\ \forall i < j$. We note $\pi(k)$ the first k samples sorted according to the descending predicted uplift $\hat{u}^{\pi}(x)$: $\pi(k) =$

[2]for completeness other metrics exist, especially if the analyst is willing to assume unconfoundedness [19]

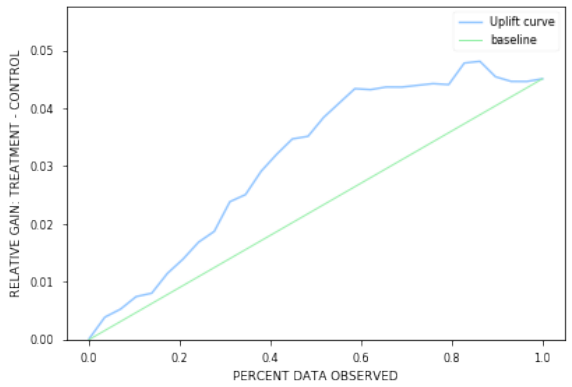

Figure 2: Typical Uplift curve, as used to compute *AUUC*.

$\{d_i \in \mathcal{D}\}_{i=1,\ldots,k}$ thus satisfying $\hat{u}(x_i) \geq \hat{u}(x_j), \forall i < j$ and $\hat{u}(x_l) \leq \hat{u}(x_i) \forall l > k, i \leq k$.

To define the uplift prediction performance let $R_\pi(k)$ be an amount of positive outcomes among the first k data points: $R_\pi(k) = \sum_{d_i \in \pi(k)} \mathbb{1}[y_i = 1]$ and we define $R^T_\pi(k)$ and $R^C_\pi(k)$ as the numbers of positive outcomes in the treatment and control groups respectively among the first k data points: $R^T_\pi(k) = R_\pi(k) \,|T = 1$, $R^C_\pi(k) = R_\pi(k) \,|T = 0$.

To define a baseline performance let also $\bar{R}^T(k)$ and $\bar{R}^C(k)$ be the numbers of positive outcomes assuming a uniform distribution of positives: $\bar{R}^T(k) = k \cdot \mathbb{E}[Y|T = 1]$, $\bar{R}^C(k) = k \cdot \mathbb{E}[Y|T = 0]$.

Finally, let $N^T_\pi(k)$ and $N^C_\pi(k)$ be the numbers of data points from treatment and control groups respectively among the first k.

Area Under Uplift Curve (*AUUC*) [7] is based on *lift curves* [20] which represent the proportion of positive outcomes (the sensitivity) as a function of the percentage of the individuals selected. Lift curve has the same ordinate as the ROC, but a different abscissa. Uplift curve is defined as the difference in lift produced by a classifier between treatment and control groups, at a particular threshold percentage k/n of all examples. Figure 2 illustrates a typical Uplift curve.

AUUC is obtained by subtracting the respective Area Under Lift (*AUL*) curves:

$$\begin{aligned} AUUC_\pi(k) &= AUL^T_\pi(k) - AUL^C_\pi(k) \\ &= \underbrace{\sum_{i=1}^{k} \left(R^T_\pi(i) - R^C_\pi(i)\right)}_{uplift} - \underbrace{\frac{k}{2}\left(\bar{R}^T(k) - \bar{R}^C(k)\right)}_{baseline} \end{aligned} \tag{5}$$

The total *AUUC* is then obtained by cumulative summation:

$$AUUC = \int_0^1 AUUC_\pi(\rho)d\rho \approx \frac{1}{n}\sum_{k=1}^{n} AUUC_\pi(k)dk \tag{6}$$

Uplift curves always start at zero and end at the difference in the total number of positive outcomes between subgroups. Higher *AUUC* indicates an overall stronger differentiation of treatment and control groups.

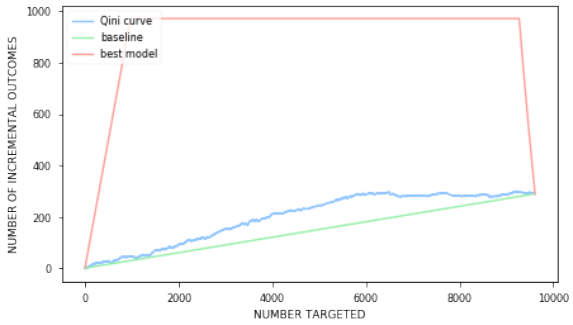

Figure 3: Typical Qini curve

Qini coefficient [1] or Q is a generalization of the Gini coefficient for the uplift prediction problem. Similarly to *AUUC* it is based on Qini curve, which shows the cumulative number of the incremental positive outcomes or uplift (vertical axis) as a function of the number of customers treated (horizontal axis). The formluation is as follows:

$$\begin{aligned} Q_\pi(k) = &\underbrace{\sum_{i=1}^{k}\left(R^T_\pi(i) - R^C_\pi(i)\frac{N^T_\pi(k)}{N^C_\pi(k)}\right)}_{uplift} \\ &- \underbrace{\frac{k}{2}\left(\bar{R}^T(k) - \bar{R}^C(k)\right)}_{baseline} \end{aligned} \tag{7}$$

A perfect model assigns higher scores to all treated individuals with positive outcomes than any individuals with negative outcomes. Thus the perfect model initially climbs at 45°, reflecting positive outcomes which are assumed to be caused by treatment. After that the graph proceeds horizontally and then climbs at 45° down due to the negative effect. In contrast, random targeting results in a diagonal line from $(0, 0)$ to (N, n) where N is the population size and n is the number of positive outcomes achieved if everyone is targeted. Real models usually fall somewhere between these two curves, forming a broadly convex curve above the diagonal, as shown on Figure 3.

Given these curves we can now define the Qini coefficient Q for binary outcomes as the ratio of the actual uplift gains curve above the diagonal to that of the optimum Qini curve:

$$Q_\pi = \frac{\sum_{k=1}^{n} Q_\pi(k)dk}{\sum_{k=1}^{n} Q_{\pi^*}(k)dk} \tag{8}$$

where π^* relates for the optimal ordering. Therefore Q theoretically lies in the range $[-1, 1]$.

Choice of metric for this task can seem unclear at first since both Equations 5 and 7 share the same high level form: a cumulative

sum of uplifts in increasing share of the population penalized by subtracting a baseline corresponding to a random model.

A first difference is that Qini corrects uplifts of selected individuals with respect to the number of individuals in treatment/control using the $N_\pi^T(k)/N_\pi^C(k)$ factor. Imagine a model selecting majorly treated individuals at a given k. The uplift part of $AUUC(k)$ can be maximized by accurately selecting positive among treated, even if there is a large proportion of positives in selected control individuals. Contrarily, $Q(k)$ would penalize such a situation. We observe in practice that Qini tend to be harder to maximize but should be preferred for model selection as it is robust to this group selection effect. Also, given that at inference time uplift models are used to predict both counter-factual outcomes we should prefer a metric that evaluates accordingly.

A second advantage of Qini is that it is normalized (8) and thus more comparable when datasets are updated over time, a typical case in some applications. We report Qini metrics in the rest of this paper.

4.2 Methods

The most basic method to predict uplift is **Two-Model** method, which uses two separate probabilistic models - first one fits on treatment group and predicts probability $P_T(Y = 1|X)$ while second one uses control group and predicts $P_C(Y = 1|X)$. Uplift then can be computed as $\hat{u}^{2m}(x) = \hat{P}_T(Y = 1|X = x) - \hat{P}_C(Y = 1|X = x)$. For this method any classification model can be used and if both of classifiers perform well, uplift model will also perform highly.

Jaskowski and Jaroszewicz [8] propose a **Class variable transformation or Revert Label** method for adapting standard classification models to the uplift case. Authors create a new label Z as follows:

$$Z = \begin{cases} 1 & \text{if } T = 1 \text{ and } Y = 1, \\ 1 & \text{if } T = 0 \text{ and } Y = 0, \\ 0 & \text{otherwise.} \end{cases}$$

and for uplift prediction in case of balanced treatment-control subgroups they obtain:

$$P_T(Y = 1|X) - P_C(Y = 1|X) = 2P(Z = 1|X) - 1.$$

As in the two-model method, any classifier can be used to predict $P(Z = 1|X)$.

Other methods include some transformed variants of SVM [13], [7] and tree-based algorithms [6], [21], [14]. SVM algorithms designed for uplift prediction have specific tasks such as construction of two separating hyperplanes instead of one or optimizing of ranking measure between pairs of examples. Tree-based methods incur finding splits in the data that optimize local variants of uplift. Both families of methods have in common that they are generally not trivial to scale in terms of either learning or inference time.

5 EXPERIMENTS

In this section we compare performance of the uplift prediction models on a Hillstrom and CRITEO-UPLIFT1 datasets. The focus is not necessarily on providing the best possible baseline but rather to highlight the fact that the Criteo dataset is a natural extension of Hillstrom, scaling up both in size and challenge while permitting to obtain statistical significance in the results. For this reason we chose Two-Model and Revert-Label approaches as they scale easily and their performance is quite close in our experience.

For the experiments we firstly preprocess datasets, specifically we normalize the features, for the classification we use Logistic Regression model with default parameters from `Scikit-Learn` [22] Python library. Then we do each experiment in the following way: we do 30 stratified random train/test splits both for treatment and control groups with a ratio 70/30, for Hillstrom dataset we use full dataset and for the CRITEO-UPLIFT1 we compare performance on 100,000, 1,000,000 and 10,000,000 randomly picked data points and on full dataset as well. For both datasets we use outcomes "visit" and "conversion".

	$Qini - 2M$	$Qini - RL$
Hillstrom-Visit-full	0.0614 ± 0.0207	0.0609 ± 0.0174
CU1-Visit-5.10^4	0.1703 ± 0.0321	0.3154 ± 0.0134
CU1-Visit-10^5	0.1500 ± 0.0246	0.3120 ± 0.0055
CU1-Visit-10^6	0.1937 ± 0.0108	0.3127 ± 0.0026
CU1-Visit-10^7	0.1910 ± 0.0027	0.3135 ± 0.0009
CU1-Visit-full	0.1872 ± 0.0024	0.3119 ± 0.0006
Hillstrom-Conversion-full	0.0914 ± 0.0804	-0.0109 ± 0.1174
CU1-Conversion-5.10^4	0.2992 ± 0.0998	0.2283 ± 0.0668
CU1-Conversion-10^5	0.3560 ± 0.0673	0.2542 ± 0.0324
CU1-Conversion-10^6	0.2331 ± 0.0261	0.2486 ± 0.0133
CU1-Conversion-10^7	0.2329 ± 0.0092	0.2536 ± 0.0051
CU1-Conversion-full	0.2315 ± 0.0049	0.2671 ± 0.0030

Table 4: Comparison of Qini scores for Two-Model ($2M$) and Revert-Label (RL) approaches on increasingly challenging datasets. Confidence is at 10% level.

Table 4 presents the results of the experiments. For readability's sake we also provide the same results in Figure 4 and 5 for visit and conversions, respectively. A first general comment is that the Qini values for the Criteo dataset are much bigger in general. For both visit and conversion targets on Hillstrom dataset the two selected methods are indistinguishable by their Qini score as their confidence intervals overlap almost entirely. For visits one begins to reach significance with the smallest extract of the Criteo dataset ($CU1 - 5e4$), which is comparable by the size with Hillstrom data. Logically, as conversions are rare and suffer from high variance one must use the largest samples $CU1 - 1e7$ or $CU1 - full$ to obtain a similar result with conversion as a target. Hence it justifies the need for a large dataset for such a challenging target.

6 CONCLUSION

We have highlighted the need for large scale benchmarks for uplift modeling in the digital advertising industry and released an open dataset two orders of magnitude larger and more challenging than previously available. We have discussed the collection and sanity checks for its counter-factual, as well as usable nature. In particular we have shown that it enables research in uplift prediction with imbalanced treatment and response levels (e.g. conversions). We

have also indicated a few other tasks for which this dataset can be useful.

Future research in uplift prediction could encompass the scaling of existing, involved methods as well as designing methods fit for imbalanced treatment level and low average response.

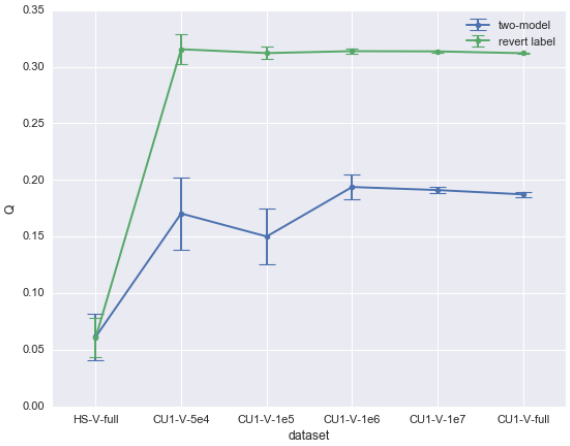

Figure 4: Comparison of Q values within different datasets using "visit" outcome.

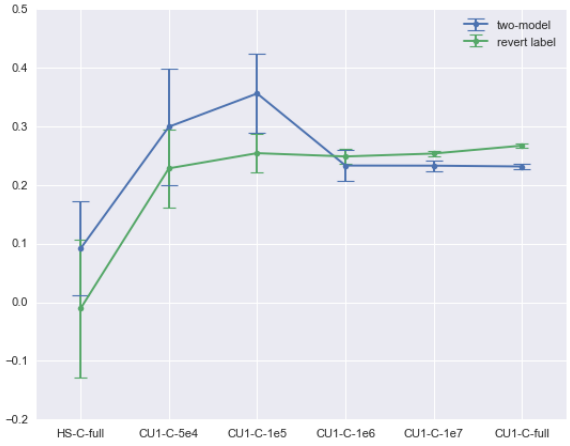

Figure 5: Comparison of Q values within different datasets using "conversion" outcome.

REFERENCES

[1] Nicholas J Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, (3):14–21, 2007.

[2] Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. Causally motivated attribution for online advertising. *Proceedings of the Sixth ...*, pages 7–9, 2012.

[3] Brett Gordon, Florian Zettelmeyer, Neha Bhargava Facebook Dan Chapsky Facebook, Gabrielle Gibbs, and Joseph Davin. A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook *. 2016.

[4] Garrett A. Johnson, Randall A. Lewis, and Elmar I. Nubbemeyer. Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness. *Journal of Marketing Research*, page jmr.15.0297, 3 2017.

[5] Ikko Yamane, Florian Yger, and Masashi Sugiyama. Uplift Modeling from Separate Labels. pages 1–18, 2018.

[6] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. pages 303–327, 2012.

[7] Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support Vector Machines for Differential Prediction. *Machine learning and knowledge discovery in databases : European Conference, ECML PKDD ... : proceedings. ECML PKDD (Conference)*, 8725:50–65, 2014.

[8] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. *ICML Workshop on Clinical Data Analysis*, 2012.

[9] Yan Zhao, Xiao Fang, and David Simchi-Levi. Uplift Modeling with Multiple Treatments and General Response Types. 2017.

[10] Kevin Hillstrom. The MineThatData e-mail analytics and data mining challenge, 2008.

[11] Piotr Rzepakowski and Szymon Jaroszewicz. Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, 2012(2):43–50, 2012.

[12] Nj Radcliffe and Pd Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic ...*, (section 6):1–33, 2011.

[13] Lukasz Zaniewicz and Szymon Jaroszewicz. Support Vector Machines for Uplift Modeling. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops*, ICDMW '13, pages 131–138, Washington, DC, USA, 2013. IEEE Computer Society.

[14] Szymon Sołtys Michałand Jaroszewicz and Piotr Rzepakowski. Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, 29(6):1531–1559, 2015.

[15] Judea. Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, 2000.

[16] Ron Berman. Beyond the Last Touch : Attribution in Online Advertising. *Preliminary Version*, 2013.

[17] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering Causal Signals in Images.

[18] Susan Athey and Guido Imbens. Recursive Partitioning for Heterogeneous Causal Effects. 4 2015.

[19] Pierre Gutierrez and Jean-Yves Gérardy. Causal Inference and Uplift Modeling A review of the literature. 67:1–13, 2016.

[20] Stéphane Tufféry. *Data mining and statistics for decision making*, volume 2. Wiley Chichester, 2011.

[21] Nicholas J Radcliffe and Patrick D Surry. Real-World Uplift Modelling with Significance-Based Uplift Trees. 2011.

[22] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830, 2011.