Uplift Prediction with Dependent Feature Representation in Imbalanced Treatment and Control Conditions

Artem Betlei[†,‡], Eustache Diemert[‡], and Massih-Reza Amini[†]

[‡]Criteo Research, [†]UGA/CNRS LIG,
Grenoble, France
{a.betlei,e.diemert}@criteo.com
Massih-Reza.Amini@univ-grenoble-alpes.fr

Abstract. Uplift prediction concerns the causal impact of a treatment over individuals and it has attracted a lot of attention in the machine learning community these past years. In this paper, we consider a typical situation where the learner has access to an imbalanced treatment and control data collection affecting the performance of the existing approaches. Inspired from transfer and multi-task learning paradigms, our approach overcomes this problem by sharing the feature representation of observations. Furthermore, we provide a unified framework for the existing evaluation metrics and discuss their merits. Our experimental results, over a large-scale collection show the benefits of the proposed approaches.

Keywords: Uplift Prediction, Causal Inference, Digital Advertising, Supervised Learning

1 Introduction

Uplift prediction is mostly studied in digital advertising and personalized medicine. In the former, the treatment is exposure to different ads [**?**] while in the latter, the treatment is usually a medication [**?**]. In both cases the aim is to predict if the treatment over an individual *would be* more preferable or not.

The ultimate goal of uplift models is to lead a policy that makes decisions over future instances. Such a decision could, for example, be to focus the advertising budget on users on whom it will be the most profitable (and possibly less annoying). One can also imagine to use such a model as a building block for a reinforcement learning algorithm that would take advantage of the predict uplift to choose relevant actions given a state. In any of these applications it is essential to enforce a causal interpretation of the uplift model in order to inform further actions.

In this paper we focus on uplift approaches that would scale to industrial applications, especially in terms of learning and inference time.

Our main contributions are threefold.

1. First, we introduce two novel approaches that tackle the case of imbalanced treatment and control datasets and discuss their merits (Section **??**)
2. we then provide a unified view of existing evaluation metrics and indicate which one should be preferred for a given application (Section **??**)
3. Finally, we evaluate the proposed approaches on a real-life collection and produce palpable evidence of their practical usefulness (Section **??**)

2 Problem formulation

The causal uplift $U(x)$ is the expected difference indicating if an individual *should* take a treatment or not. We formalize it using Pearl's causal inference framework [**?**] as

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

Conversely, the conditional uplift $u(x)$ in Equation **??** 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:

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

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 outcome coming from a natural experiment. But 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 $U(x) \equiv u(x)$. Therefore we assume a dataset composed of i.i.d. samples of the joint covariates X, label Y and treatment T variables:

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

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, if the subject is treated or not, respectively.

3 Proposed approaches

In this section, we present two uplift prediction methods for large-scale data case which attempt to take advantage of the relatedness of response in treatment and control groups during the learning process. We posit that this general idea is useful when the amount of data varies drastically between groups.

Dependent Data Representation (DDR) approach is based on a Classifier Chains method [**?**] originally developed for multi-label classification problems. The idea is that if there are L different labels, one can build L different classifiers, each of which solves the problem of binary classification and at the training process each next classifier uses predictions of the previous ones as extra features.

We use the same idea for our problem in two steps. At the beginning we train a first classifier on control data:

$$P_C = P(Y = 1|X = x, T = 0),$$

then we use predictions P_C as an extra feature for the classifier learning on the treatment data, effectively injecting a dependency between the two datasets:

$$P_T = P(Y = 1|X = x, \hat{P}_C(x) = p, T = 1).$$

To obtain uplift for each individual we compute the difference:

$$\hat{u}^{DDR}(x) = \hat{P}_T(x, \hat{P}_C(x)) - \hat{P}_C(x)$$

Intuitively, the second classifier is learning the difference between the expected outcome in treatment and control, that is the uplift itself. Examination of the weights of this uplift classifier could also lead to interesting information on the role of different features in explaining the treatment outcome.

Shared Data Representation (SDR) approach for uplift prediction is based on a popular implementation of the multi-task framework [**?**]. A predictor is learned on a modified features representation that allows to learn related tasks jointly and with a single loss. We specialize this approach considering predicting outcomes in control and treatment groups as the related tasks.

The general form of the model is given by

$$P(Y|T = t, X = x) = f(\langle w_0, x\rangle + \mathbb{1}_{[t=1]}\langle w_t, x\rangle + \mathbb{1}_{[c=1]}\langle w_c, x\rangle) \tag{3}$$

with f an arbitrary link function. Practically speaking we augment the dataset by stacking the original features with a conjunction of the treatment group indicator and the same features. Letting $\mathbf{D}_T$ and $\mathbf{D}_C$ be the covariates from treatment and control groups respectively such that $\mathbf{D}_T \cup \mathbf{D}_C = \mathbf{D}$ we obtain the following shared learning representation:

$$\mathbf{D}^{SDR}_{train} = \begin{bmatrix} \mathbf{D}_T & \mathbf{D}_T & 0 \\ \mathbf{D}_C & 0 & \mathbf{D}_C \end{bmatrix}$$

So a single vector of weights $\mathbf{w}$ is learned jointly as $\mathbf{w} = [\mathbf{w}_0\ \mathbf{w}_T\ \mathbf{w}_C]$ where $\mathbf{w}_0$ is a vector of weights that relate to the original features and $\mathbf{w}_T$ and $\mathbf{w}_C$ are corresponding to treatment/control conjunction features.

At inference time we compute the difference between predicted probabilities using two representations of the individual features, corresponding to the counterfactual outcomes:

$$\hat{u}^{SDR}(x) = \hat{P}(Y = 1|\left[x\ x\ 0\right]) - \hat{P}(Y = 1|\left[x\ 0\ x\right])$$

An advantage of this method is the possibility to assign different regularization penalties for $\mathbf{w}_0$ (λ_0) and $\mathbf{w}_T$ / $\mathbf{w}_C$ (λ_1). In this way it is possible to control the strength of the connection between the tasks. As reported by Chapelle, it is equivalent to rescaling the conjunction features by $\sqrt{\lambda_0/\lambda_1}$. Intuitively this model allows to learn a common set of weights for predicting the global, average outcome whilst keeping enough capacity to express the peculiar influence of features in the treatment or control conditions.

4 Related work

We review existing uplift prediction methods first to highlight links and differences with the proposed methods and will then proceed to the evaluation metrics.

4.1 Learning Approches

Here we describe current methods for uplift prediction and explain the advantages and drawbacks of them.

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. At the same time the main goal of the models is to predict outcomes separately but not exactly the uplift. In cases where the average response is low and/or noisy there is a risk that the difference of predictions would be very noisy too.

DDR can be seen as an extension of the two-model approach, the difference in interpretation is that adding an extra feature to the classifier learned on treatment group we add a knowledge about control group, thus we learn directly to transfer uplift to the unobserved, counter-factual case.

Jaskowski and Jaroszewicz [**?**] 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: $Z = YT + (1 - Y)(1 - T)$, 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 $P(Z = 1|X)$. New label Z unites two subgroups with different outcomes, so it is not clear how difficult it would be for a model to find optimal weights, especially on the imbalanced outcomes case. Another drawback is that all capacity of a model is spent for direct uplift prediction, without any assumptions connected with treatment and control subgroups.

Other methods include some transformed variants of SVM [**?**], [**?**] and tree-based algorithms [**?**], [**?**], [**?**]. 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.

4.2 Metrics

We now describe the two major uplift metrics. As both are based on an ordering of samples according to their predicted uplift scores we assign following 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 $\hat{u}^{\pi}(x)$:

$$\pi(k) = \{d_i \in \mathcal{D}\}_{i=1,\dots,k}$$

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 ($AUUC$) [**?**] is based on the *lift curves* [**?**] which represent the proportion of positive outcomes (the sensitivity) as a function of the percentage of the individuals selected. 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.

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

$$AUUC_{\pi}(k) = AUL^T_{\pi}(k) - AUL^C_{\pi}(k) = \underbrace{\sum_{i=1}^{k} \Big(R^T_{\pi}(i) - R^C_{\pi}(i)\Big)}_{uplift} - \underbrace{\frac{k}{2}\Big(\bar{R}^T(k) - \bar{R}^C(k)\Big)}_{baseline} \tag{4}$$

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

$$AUUC = \int_0^1 AUUC_\pi(\rho)d\rho \approx \frac{1}{n}\sum_{k=1}^{n} AUUC_\pi(k)dk \tag{5}$$

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.

Qini coefficient [**?**] 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:

$$Q_\pi(k) = \underbrace{\sum_{i=1}^{k}\left(R_\pi^T(i) - R_\pi^C(i)\frac{N_\pi^T(k)}{N_\pi^C(k)}\right)}_{uplift} - \underbrace{\frac{k}{2}\left(\bar{R}^T(k) - \bar{R}^C(k)\right)}_{baseline} \tag{6}$$

A perfect model assigns higher scores to all treated individuals with positive outcomes than any individuals with negative outcomes. Thus at the beginning perfect model 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. 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:

$$Q_\pi = \frac{\sum_{k=1}^{n} Q_\pi(k)dk}{\sum_{k=1}^{n} Q_{\pi^*}(k)dk} \tag{7}$$

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 **??** and **??** 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 (**??**) 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.

5 Experiments

In this section we define a benchmark for the experiments, present a comparison between proposed and other uplift prediction methods.

5.1 Benchmark

It is difficult to obtain data for learning an unbiased uplift prediction model (i.e. data from random treatment assignment). We only know of two unbiased, large scale datasets. **Hillstrom dataset** [**?**] contains results of an e-mail campaign for an Internet based retailer. The dataset contains information about 64,000 customers involved in an e-mail test who were randomly chosen to receive men's, women's merchandise e-mail campaigns or not receive an e-mail. We use the no-email vs women e-mail split with "visit" as outcome as in [**?**]. Our second dataset is **CRITEO-UPLIFT1**[1] which is constructed by assembling data resulting from incrementality tests, a particular randomized trial procedure where a random part of the population is prevented from being targeted by advertising. It consists of 25M rows, each one representing a user with 12 features, a treatment indicator and 2 labels (visits and conversions).

For the experiments we firstly preprocess datasets, specifically we binarize categorical variables and normalize the features, for the classification we use Logistic Regression model from `Scikit-Learn` [**?**] Python library as it has fast learning and inference processes. Then we do each experiment in the following way: we do 50 stratified random train/test splits both for treatment and control groups with a ratio 70/30, during learning process we tune parameters of each model on a grid search. For DDR and SDR we use the regularization trick that we explained earlier, we tune additional regularization terms on a grid search as well. To check statistical significance we use two-sample paired t-test at 5% confidence level (marked in bold in the tables when positive).

5.2 Performance of Dependent Data Representation

We compare DDR with a Two-Model as first is an extention of the second, results are shown on Table **??**. We use Hillstrom dataset with a "visit" outcome

[1] this dataset will be released shortly at `http://research.criteo.com/outreach`

and cover three cases: firstly we compare approaches on a full dataset, then reduce control group randomly choosing 50% of it and for the last experiment we randomly choose 10% of control group to check how methods will perform with imbalanced data case. Indeed it is usually the case that the control group is kept to a minimum share so as not to hurt global treatment efficiency (e.g. ad revenue). As we can see, DDR significantly outperform Two-Model on imbalanced cases.

	BALANCED T/C	IMBALANCED T/C (50% OF C GROUP)	HIGHLY IMBALANCED T/C (10% OF C GROUP)
TWO MODEL	0.06856	0.06292	0.03979
DDR	0.06866	**0.06444**	**0.04557**

Table 1. Performances of Two-Model and DDR approaches measured as mean Q.

Different directions of DDR As DDR approach is based on a consecutive learning of two classifiers, there are two ways of learning - to fit first model on treatment group and then use output as a feature for the second one and fit it on a control part (we denote it as $T \rightarrow C$), or vice versa ($C \rightarrow T$). Table **??** indicates that both approaches are comparable in the balanced case but $C \rightarrow T$ direction is preferable in other cases (at least with this dataset). Since the test set has more treated examples it makes sense that the stronger predictor obtained on this group by using information from predicted uplift on control performs best.

	BALANCED T/C	IMBALANCED T/C (50% OF C GROUP)	HIGHLY IMBALANCED T/C (10% OF C GROUP)
DDR ($T \rightarrow C$)	0.06895	0.06394	0.03979
DDR ($C \rightarrow T$)	0.06866	0.06444	0.04557

Table 2. Comparison of directions of learning in DDR approach (Q).

Complexity of treatment effect with DDR

To investigate complexity of the link between treatment and control group we use a dummy classifier (predicting the average within-group response) successively for one of treatment or control group while still using the regular model for the remaining group. Intuitively if the treatment effect is a constant, additive uplift then a simple re-calibration using a dummy model should be good enough. Conversely if there is a rich interaction between feature and treatment to explain outcome a second, a dummy classifier would perform poorly. Table **??** indicates that the rich interaction hypothesis seems more plausible in this case, with maybe an even richer one in treated case.

5.3 Performance of Shared Data Representation

Here we compare SDR approach with Revert Label because of a similar nature of the uplift prediction. Revert Label model is learned with samples reweighting

	BALANCED T/C
DDR	0.06866
DDR (DUMMY FOR C GROUP)	0.04246
DDR (DUMMY FOR T GROUP)	0.01712

Table 3. Comparison between different variants of DDR approach.

as in the original paper. Table **??** indicates that SDR significantly outperforms Revert Label on imbalanced cases. Note that due to heavy down-sampling in the imbalanced cases it is not trivial to compare Q values between columns.

	BALANCED T/C	IMBALANCED T/C (50% OF C GROUP)	HIGHLY IMBALANCED T/C (10% OF C GROUP)
REVERT LABEL	0.06879	0.06450	0.05518
SDR	0.06967	**0.06945**	**0.08842**

Table 4. Performances of Revert Label and SDR approaches measured as mean Q.

Usefullness of conjunction features

In order to check usefulness of conjunctions features with SDR we compare it with a trivial variant in which we simply add an indicator variable for treatment instead of the whole feature set. This allows the model to learn only a simple re-calibration of the prediction for treated/control. Table **??** indicates that it strongly degrades model performance.

	SDR (STANDARD)	SDR (T/C INDICATOR)
Q	**0.06967**	0.02706

Table 5. Comparison between variants of SDR in balanced treatment/control conditions.

Performance in imbalanced outcome condition

We also compare SDR approach with Revert Label on CRITEO-UPLIFT1 dataset with conversion as outcome on a random sample of 50,000. Ratio between C and T group is 0.18 so it is highly imbalanced case as well but the outcome is also imbalanced with average level at only .00229. Table **??** indicates that SDR again significantly outperforms Revert Label in this setting.

6 Conclusion

We proposed two new approaches for the Uplift Prediction problem based on dependent and shared data representations. Experiments show that they outperform current methods in imbalanced treatment conditions. In particular they

	Revert Label	SDR
Q	0.25680	**0.54228**

Table 6. Performances of Revert Label and SDR in highly imbalanced conditions for both treatment and outcome.

allow to learn rich interaction between the features and treatment to explain response. Future research would include learning more complex (highly non-linear) data representations permitting even richer interactions between features and treatment.