# Uplift Modeling with ROC: An SRL Case Study

Houssam Nassif, Finn Kuusisto, Elizabeth S. Burnside, and Jude Shavlik

University of Wisconsin, Madison, USA

**Abstract.** Uplift modeling is a classification method that determines the incremental impact of an action on a given population. Uplift modeling aims at maximizing the area under the uplift curve, which is the difference between the subject and control sets' area under the lift curve. Lift and uplift curves are seldom used outside of the marketing domain, whereas the related ROC curve is frequently used in multiple areas. Achieving a good uplift using an ROC-based model instead of lift may be more intuitive in several areas, and may help uplift modeling reach a wider audience.

We alter SAYL, an uplift-modeling statistical relational learner, to use ROC instead of lift. We test our approach on a screening mammography dataset. SAYL-ROC outperforms SAYL on our data, though not significantly, suggesting that ROC can be used for uplift modeling. On the other hand, SAYL-ROC returns larger models, reducing interpretability.

## 1 Introduction

*Uplift modeling* is a modeling and classification method initially used in marketing to determine the incremental impact of an advertising campaign on a given population [8]. Seminal work includes Radcliffe and Surry's true response modeling [8], Lo's true lift model [4], and Hansotia and Rukstales' incremental value modeling [3]. In some applications, especially medical decision support systems, gaining insight into the underlying classification logic can be as important as system performance. Reviewing the classification logic in medical problems can be an important method to discover disease patterns that may not be known or easily otherwise gleaned from the data. Such insight can be achieved using rule-learners. Decision trees [9, 10], inductive logic programming (ILP) [7], and statistical relational learning (SRL) [6] methods have been proposed.

Uplift modeling aims at maximizing uplift, which is the difference in a model or intervention $M$'s lift scores over the subject and control sets:

$$Uplift_M = Lift_M(subject) - Lift_M(control). \tag{1}$$

Given a fraction $\rho$ such that $0 \leq \rho \leq 1$, a model $M$'s lift is defined as the number of positive examples amongst the model's $\rho$-highest ranking examples. Uplift thus captures the additional number of positive examples obtained due to the intervention. Quality of an uplift model is often evaluated by computing an uplift curve [9], generated by ranging $\rho$ from 0 to 1 and plotting $Uplift_M$. The higher the uplift curve, the more profitable a marketing model/intervention is. The area under the uplift curve (AUU) is often used as a metric to optimize.

Let $P$ be the number of positive examples and $N$ the number of negative examples in a given dataset $D$. Lift represents the number of true positives detected by model $m$ amongst the top-ranked fraction $\rho$. Varying $\rho \in [0,1]$ produces a lift curve. The area under the lift curve (AUL) for a given model and data becomes:

$$AUL = \int Lift(D,\rho)d\rho \approx \frac{1}{2}\sum_{k=1}^{P+N}(\rho_{k+1}-\rho_k)(Lift(D,\rho_{k+1})+Lift(D,\rho_k)) \quad (2)$$

Let $s$ be the subject set, and $c$ the controls. For a given $\rho$, we can rewrite equation 1 as:

$$Uplift_M(\rho) = Lift_M(s,\rho) - Lift_M(c,\rho). \quad (3)$$

Since uplift is a function of a single value for $\rho$, the area under the uplift curve (AUU) is the difference between the areas under the lift curves (AUL) for the subjects and the controls, $\Delta(AUL)$:

$$AUU = AUL_s - AUL_c = \Delta(AUL). \quad (4)$$

Lift and uplift curves are seldom used outside of the marketing domain, whereas the related ROC curve is frequently used in the machine learning and biomedical informatics communities. Especially in the biomedical domain, using ROC may be more intuitive, and may help uplift modeling reach a wider audience. This work investigates the use of the area under the ROC curve (AUR) as an alternate scoring method, while still resulting in a good model uplift. We alter SAYL [6], the state-of-the-art relational uplift modeling algorithm, to select rules that optimize $\Delta(AUR)$ instead of $\Delta(AUL)$. We test our approach on a screening mammography dataset.

## 2 Lift and ROC Area Under the Curve

There is a strong connection between AUL and AUR. Let $\pi = \frac{P}{P+N}$ be the prior probability for the positive class or skew, then:

$$AUL = P * (\frac{\pi}{2} + (1-\pi)\ AUR)\ [11, p.549]. \quad (5)$$

Uplift modeling aims at optimizing uplift, the difference in lift over two sets. It constructs a new classifier such that:

$$\Delta(AUL^*) > \Delta(AUL) \quad (6)$$

As discussed in [6], by expanding and simplifying we get:

$$AUL_s^* - AUL_c^* > AUL_s - AUL_c$$

$$P_s(\frac{\pi_s}{2} + (1-\pi_s)AUR_s^*) - P_c(\frac{\pi_c}{2} - (1-\pi_c)AUR_c^*) > P_s(\frac{\pi_s}{2} + (1-\pi_s)AUR_s) - P_c(\frac{\pi_c}{2} - (1-\pi_c)AUR_c)$$

$$P_s(1-\pi_s)AUR_s^* - P_c(1-\pi_c)AUR_c^* > P_s(1-\pi_s)AUR_s - P_c(1-\pi_c)AUR_c$$

$$P_s(1-\pi_s)(AUR_s^* - AUR_s) > P_s(1-\pi_s)(AUR_c^* - AUR_c)$$

and finally

$$\frac{AUR_s^* - AUR_s}{AUR_c^* - AUR_c} > \frac{P_c}{P_s}\frac{1-\pi_c}{1-\pi_s}. \tag{7}$$

In a balanced dataset, we have $\pi_c = \pi_s = \frac{1}{2}$ and $P_c = P_s$, so we have that $\frac{P_c}{P_s}\frac{1-\pi_c}{1-\pi_s} = 1$. If the subject and control sets have the same numbers and skew, we can conclude that $\Delta(AUL^*) > \Delta(AUL)$ implies $\Delta(AUR^*) > \Delta(AUR)$. If the two sets are skewed or their numbers differ, we cannot guarantee that $\Delta(AUL^*) > \Delta(AUL)$ implies $\Delta(AUR^*) > \Delta(AUR)$, as we can increase uplift with rules that have similar accuracy but cover more cases in the positive set. In general, the two metrics are related, with uplift being more sensitive to variations in coverage when the two groups have different size.

## 3 SAYL-ROC

SAYL [6] is a Statistical Relational Learner based on SAYU [1] that integrates uplift modeling with the search for relational rules. Similar to SAYU, every valid rule generated is used to construct a Bayesian network (alongside with current theory rules) via propositionalization, but instead of constructing a single classifier, SAYL constructs two TAN [2] classifiers; one Bayes net for each of the subject and control groups. Both classifiers use the same set of attributes, but are trained only on examples from their respective groups. SAYL uses the TAN generated probabilities to construct the lift and uplift curves. If a rule improves AUU by threshold $\theta$, the rule is added to the attribute set. Otherwise, SAYL continues the search.

**Algorithm 1** SAYL

```
Rs ← {}; M_0^s, M_0^c ← InitClassifiers(Rs)
while DoSearch() do
    e_s^+ ← RandomSeed();
    ⊥_{e_s^+} ← saturate(e);
    while c ← reduce(⊥_{e_s^+}) do
        M^s, M^c ← LearnClassifiers(Rs ∪ {c});
        if Better(M^s, M^c, M_0^s, M_0^c) then
            Rs ← Rs ∪ {c}; M_0^s, M_0^c ← M^s, M^c;
            break
        end if
    end while
end while
```

The SAYL algorithm is shown as Algorithm 1. SAYL maintains a current set of clauses, $Rs$, and current reference classifiers for the subjects $M^s$ and controls $M^c$. SAYL requires separate training and tuning sets, accepting a rule only when it improves the score on both sets. This requirement is extended with the threshold of improvement $\theta$, and a minimal rule coverage requirement *minpos*. Finally, SAYL has two search modes, greedy and exploration. Refer to [6] for details.

SAYL guides the rule search by using the AUU score. It computes AUU by computing AUL for each of the groups using the two classifiers, and returning the difference $\Delta(AUL)$ (Equation 4). We implement SAYL-ROC, a SAYL variant that computes AUR instead for each of the groups using the two classifiers, and returns $\Delta(AUR)$ as a rule score to guide the search. SAYL thus optimizes for $\Delta(AUL)$, while SAYL-ROC optimizes for $\Delta(AUR)$.

## 4 Experimental Results

We test SAYL-ROC on a breast cancer mammography dataset, fully described in [5]. Our subject and control sets are respectively older and younger patients with confirmed breast cancer. Positive instances have in situ cancer, and negative instances have invasive cancer. The aim is to maximize the in situ cases' uplift.

The older cohort has 132 in situ and 401 invasive cases, while the younger one has 110 in situ and 264 invasive. The skews are $P_s = 132$, $\pi_s = \frac{132}{132+401}$ (older), and $P_c = 110$, $\pi_c = \frac{110}{110+264}$ (younger). Thus equation 7 becomes:

$$\frac{AUR_s^* - AUR_s}{AUR_c^* - AUR_c} > 0.86. \tag{8}$$

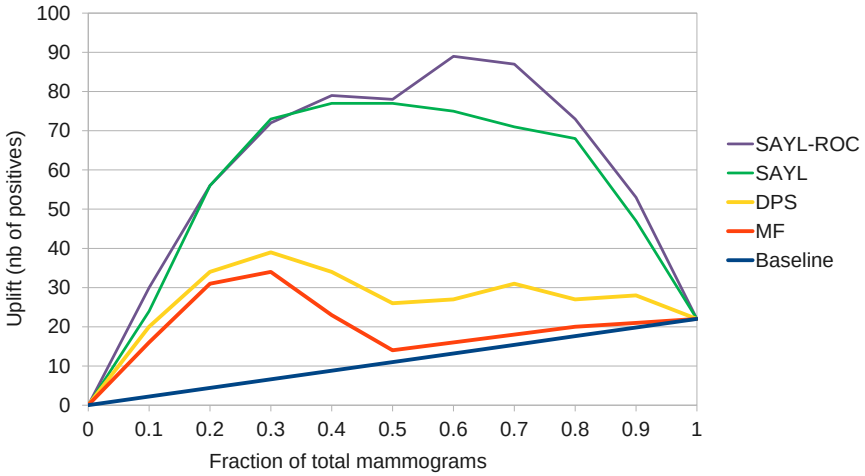
We use 10-fold cross-validation, making sure all records pertaining to the same patient are in the same fold. We run both SAYL and SAYL-ROC with a time limit of one hour per fold. For each cross-validated run, we use 4 training, 5 tuning and 1 testing folds. For each fold, we used the best combination of parameters according to a 9-fold internal cross-validation using 4 training, 4 tuning and 1 testing folds. We try both search modes, vary *minpos* between 7 and 13 (respectively 5% and 10% of older in situ examples), and set $\theta$ to 1%, 5% and 10%. We evaluate the final SAYL and SAYL-ROC models using their final uplift curves, concatenated from the results of each testing set.

**Table 1.** 10-fold cross-validated SAYL-ROC and SAYL performance. Rule number averaged over the 10 folds of theories. For comparison, we include results of Differential Prediction Search (DPS) and Model Filtering (MF) methods [7]. We compute the *p*-value comparing each method to SAYL, * indicating significance.

| Algorithm | AUU | $AUL_s$ | $AUL_c$ | Rules Avg | *p*-value |
|---|---|---|---|---|---|
| **SAYL-ROC** | 62.99 | 95.64 | 32.65 | 24.7 | 0.4316 |
| **SAYL** | 58.10 | 97.24 | 39.15 | 9.3 | - |
| **DPS** | 27.83 | 101.01 | 73.17 | 37.1 | 0.0020 * |
| **MF** | 20.90 | 100.89 | 80.99 | 19.9 | 0.0020 * |
| **Baseline** | 11.00 | 66.00 | 55.00 | - | 0.0020 * |

Table 1 compares SAYL-ROC, SAYL, and the ILP-based methods Differential Prediction Search (DPS) and Model Filtering (MF) [7], both of which had *minpos* = 13 (10% of older in situ). A baseline random classifier achieves an AUU of 11. We use the paired Mann-Whitney test at the 95% confidence level to compare two sets of experiments. We plot the uplift curves in Figure 1.

**Fig. 1.** Uplift curves for SAYL-ROC, SAYL, ILP methods Differential Prediction Search (DPS) and Model Filtering (MF), both with *minpos* = 13 [7], and baseline random classifier. Uplift curves start at 0 and end at 22, the difference between older (132) and younger (110) total in situ cases. The higher the curve, the better the uplift.




## 5 Discussion and Future Work

SAYL and SAYL-ROC significantly outperform previous methods (Table 1, Figure 1), but there is no significant difference between the two. Even though SAYL-ROC is optimizing for $\Delta(AUR)$ during its training phase, it returns a slightly better testing $\Delta(AUL)$ than SAYL, which optimizes for $\Delta(AUL)$.

This result suggests that, on a moderately subject/control skewed data, AUR can indeed be used for uplift modeling. ROC is more frequently used than lift, and may be more intuitive in many domains. Nevertheless, more experiments are needed to establish ROC-based uplift performance. We plan on measuring $\Delta(AUL)$ vs. $\Delta(AUR)$ for various Equation 7 skews.

SAYL-ROC produces as many rules as ILP-based methods, more than twice that of SAYL. The ILP theory is a collection of independent rules that each individually increases uplift [7]. It is thus easy to interpret the final model. SAYL and SAYL-ROC theory rules are conditioned on each other as nodes in a Bayesian network, decreasing rule interpretability especially in larger graphs. Individual rules may not increase uplift, but the final network does. At an average of 9.3 rules, a SAYL model is interpretable, whereas at 24.7, SAYL-ROC sacrifices interpretability.

We note that Equation 7 depends on both the positive number and skew. Even if the subject and control positive skews were equal, say $P_c = 100$, $N_c = 200$, $P_s = 10$ and $N_s = 20$, we will have $\frac{1-\pi_c}{1-\pi_s} = 1$ but $\frac{P_c}{P_s} = 10$, maintaining a subject/control Equation 7 skew.

This work uses the definition of lift as the *number* of positives amongst the $\rho$-highest ranking examples. An alternative lift definition is the *fraction* of positives amongst the $\rho$-highest ranking examples. Equation 7 then becomes:

$$\frac{AUR_s^* - AUR_s}{AUR_c^* - AUR_c} > \frac{1-\pi_c}{1-\pi_s}, \tag{9}$$

eliminating the dependence on the number of positive instances. We plan on investigating how $\Delta(AUL)$ and $\Delta(AUR)$ empirically relate under this definition.

In conclusion, SAYL-ROC exhibits a similar performance to SAYL on our data, suggesting that ROC can be used for uplift modeling. SAYL-ROC returns larger models, reducing interpretability. More experiments are needed to test ROC-based uplift over different subject/control skews.

**Acknowledgments** We thank NIH grant R01-CA165229, the Carbone Cancer Center, and NCI grant P30CA014520 for support.

## References

1. Davis, J., Burnside, E.S., de Castro Dutra, I., Page, D., Santos Costa, V.: An integrated approach to learning Bayesian Networks of rules. In: Proceedings of the 16th European Conference on Machine Learning. pp. 84–95. Porto, Portugal (2005)
2. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning 29, 131–163 (1997)
3. Hansotia, B., Rukstales, B.: Incremental value modeling. Journal of Interactive Marketing 16(3), 35–46 (2002)
4. Lo, V.S.: The true lift model - a novel data mining approach to response modeling in database marketing. SIGKDD Explorations 4(2), 78–86 (2002)
5. Nassif, H., Page, D., Ayvaci, M., Shavlik, J., Burnside, E.S.: Uncovering age-specific invasive and DCIS breast cancer rules using Inductive Logic Programming. In: ACM International Health Informatics Symposium (IHI). pp. 76–82. Arlington, VA (2010)
6. Nassif, H., Kuusisto, F., Burnside, E.S., Page, D., Shavlik, J., Santos Costa, V.: Score as you lift (SAYL): A statistical relational learning approach to uplift modeling. In: European Conference on Machine Learning (ECML-PKDD). pp. 595–611. Prague (2013)
7. Nassif, H., Santos Costa, V., Burnside, E.S., Page, D.: Relational differential prediction. In: European Conference on Machine Learning (ECML-PKDD). pp. 617–632. Bristol, UK (2012)
8. Radcliffe, N.J., Surry, P.D.: Differential response analysis: Modeling true response by isolating the effect of a single action. In: Credit Scoring and Credit Control VI. Edinburgh, Scotland (1999)
9. Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions (2011)
10. Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems 32, 303–327 (2012)
11. Tufféry, S.: Data Mining and Statistics for Decision Making. John Wiley & Sons, 2nd edn. (2011)