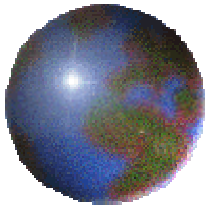
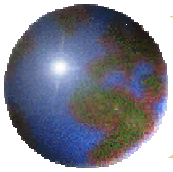


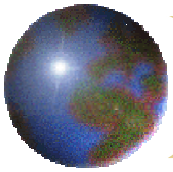
# ブースティング手法による顧客スコアリング

筑波大学システム情報工学研究科
社会システム工学専攻
佐野　夏樹



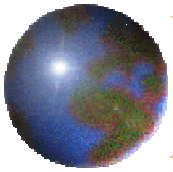
# 発表内容

- 本研究の概要
- ブースティングとは
- 優良顧客のスコアリングとは
- AdaBoostのスコアリングへの適用
- 実証分析
- 結論



# ブースティング

- たくさんの学習機械を組み合わせて統合学習機械を作ることで，もともとの学習機械の精度(accuracy)を向上させる手法．代表的なものにAdaBoost (Freund and Schapire)などがある．



# *AdaBoost M1* (Freund & Schapire 1997)

1.重みの初期値を $w_i=1/n,\ i = 1,2,\ldots,n$ とする.

2. t=1 to T:

(a)重み$w_i$ を用いて学習データに機械 $f_t(x)$をあてはめる.

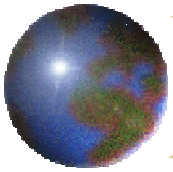
(b)重み付き誤り率 $\varepsilon_t = \sum_{i=1}^{n} w_i I(y \neq f_t(x_i))$

信頼度 $\beta_t = \log\left((1-\varepsilon_t)/\varepsilon_t\right)$ を求める.

(c) $w_i \leftarrow w_i \cdot \exp\left[\beta_t \cdot I(y_i \neq f_t(x_i))\right], i = 1,2,\ldots,n$

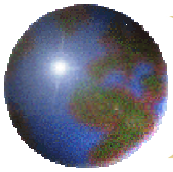
重みを更新し, $\sum_i w_i = 1$ となるように正規化する.

3.識別機械 $\mathrm{sign}(\mathrm{F_T}(x)) = \mathrm{sign}\left(\sum_{t=1}^{T} \beta_t f_t(x)\right)$ とする.



# 優良顧客のスコアリング

- スコアリングとは顧客を優良顧客から順に数値化すること．優良顧客に絞ったマーケティングキャンペーンを展開することで顧客のロイヤルティを上げたり，経費の削減を行うことができる．
- 先行研究
  - 竹林実，佐野夏樹，鈴木秀男「AdaBoostによる顧客スコアリング」，日本オペレーションズ・リサーチ学会２００３年秋季研究発表会
  - 後藤 正輝，村山 一穂，門間 公志，香田 正人「データマイニング手法によるスコアリングモデルの開発」，「Direct Marketing Review」，vol.1, 19-32. (2002)

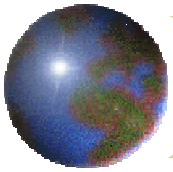


# *AdaBoost*によるスコアリングモデル

$$F(x) = \sum_{t=1}^{T} \beta_t f_t(x)$$

**による顧客ランクを提案**

**これは次の事後確率推定に基づいている．**



# AdaBoostによる事後確率推定

*Friedman, Hastie and Tibshirani(1998)*

$$\Pr\left(y=+1\middle|x\right)=\frac{\exp\left(F(x)\right)}{\exp\left(F(x)\right)+\exp\left(-F(x)\right)}$$

**ここで左辺は当該顧客が購入者となる確率であり，**

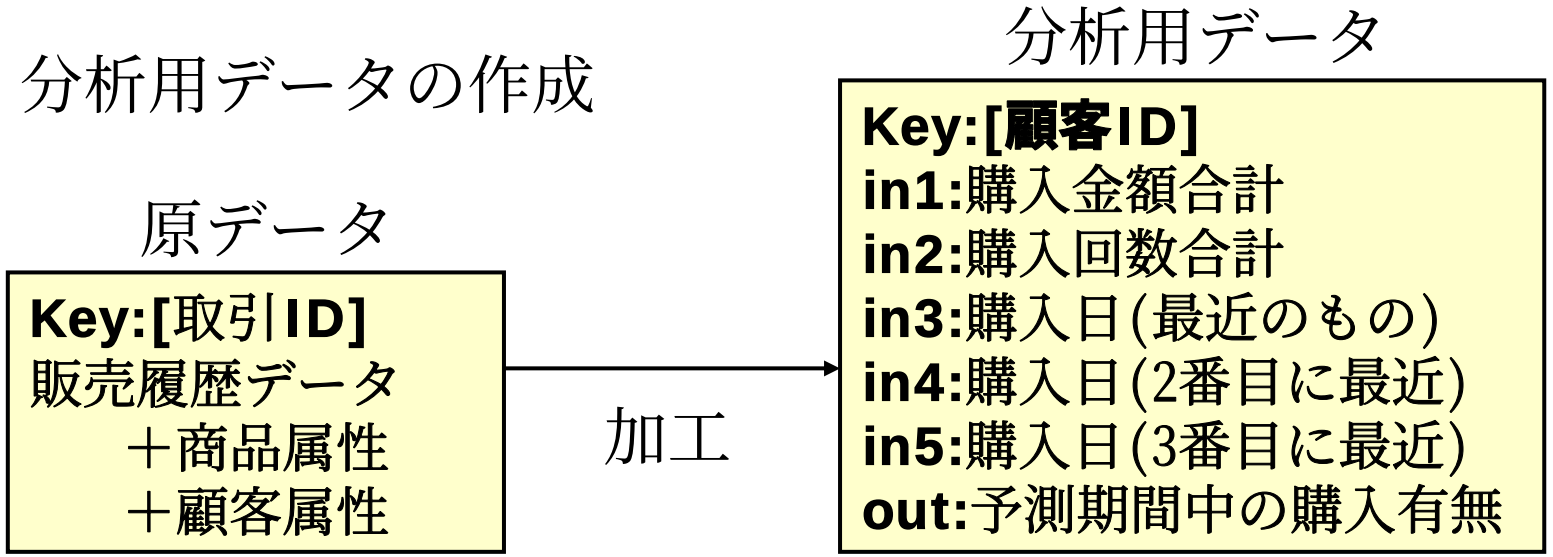
$F(x)=\sum_{t=1}^{T}\beta_t f_t(x)$ **は**AdaBoost**による出力を表す．**

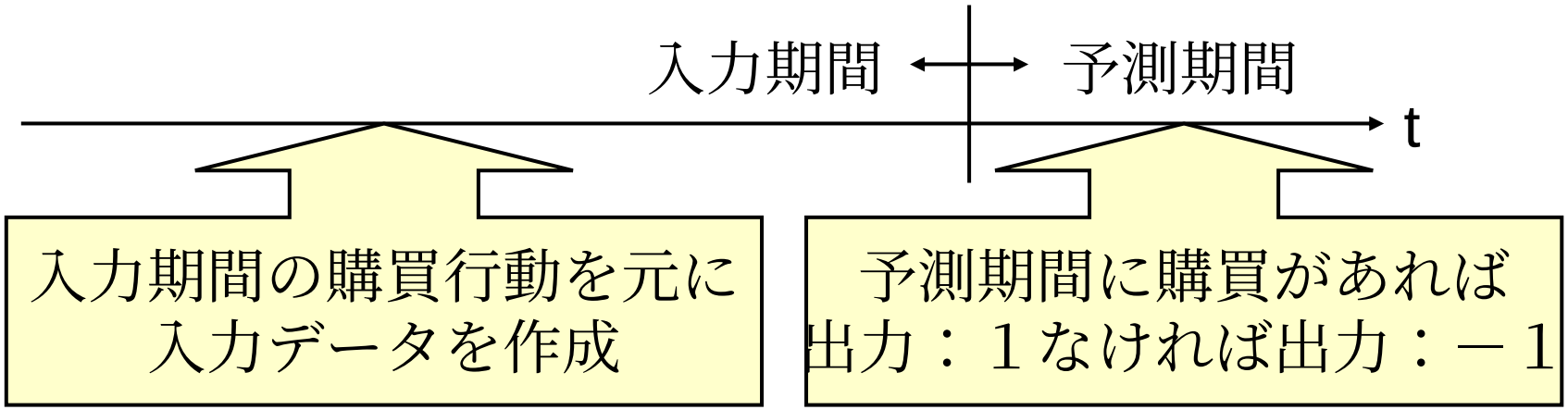
# 実証分析（1）

- ある衣料雑貨販売会社の通信販売履歴データについての顧客スコアリングモデルを作成（約12000名、3年間の履歴）.
- 基本的には購買予測にもっとも影響を与えると言われているRecently，Monetary，Frequencyに基づいてモデルを構築.
- ゲインチャートによる従来のモデルとの比較.

# データ前処理

- 分析用データの作成



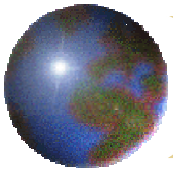



# 学習データ，テストデータにおける購入者割合

| | 学習用データ | | | テスト用データ | | |
|---|---|---|---|---|---|---|
| | 購入者数 | 総顧客数 | 購入者割合 | 購入者数 | 総顧客数 | 購入者割合 |
| 分析1 | 112 | 1056 | 10.6% | 920 | 9504 | 9.7% |
| 分析2 | 516 | 1720 | 30.0% | 516 | 5280 | 9.8% |

分析１：オリジナルデータの購入者割合（10.6%）で学習

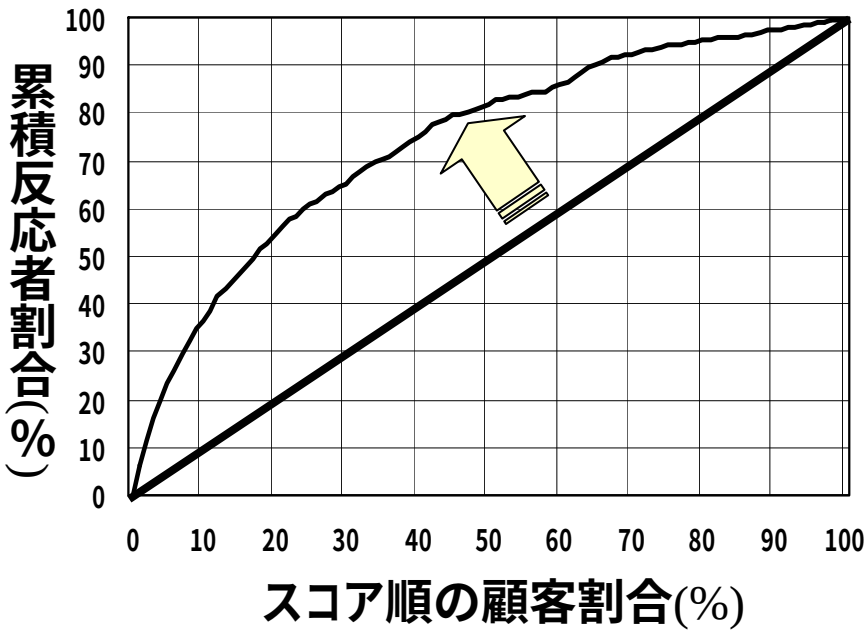
分析２：購入者割合を変更（30.0%)したデータで学習



# 実証分析（2）

- 10560人のデータ中の購入顧客の数が1032人と少ないので学習データ中の購入者割合を30%に変更した分析(分析2)を行った.
- 学習データによってモデルを作成し，テストデータに適用.
- AdaBoostの基本学習機械は深さ1の決定木とし，繰り返し数T=500とした.

# モデルの精度評価基準




累積ゲイン図 (Berry and Linoff, 2002)

予測全体のあてあまりを評価するグラフ.

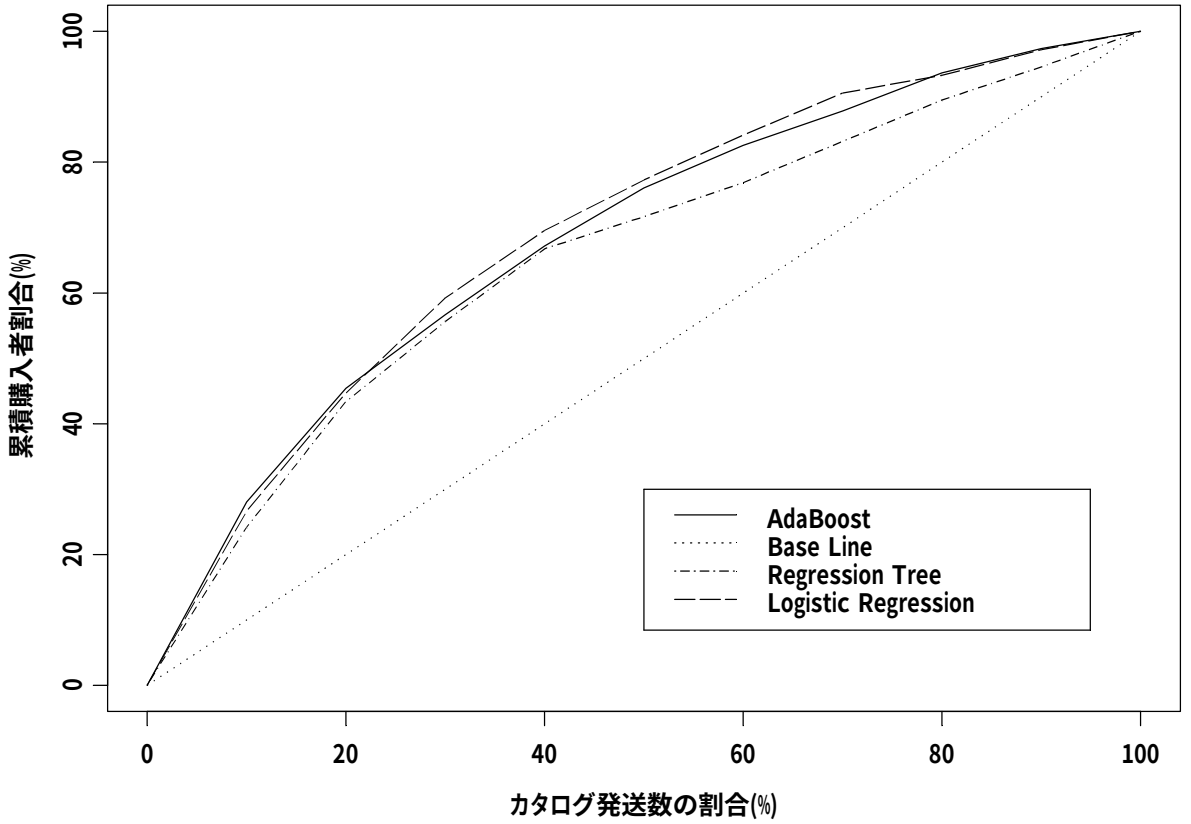
１．顧客を購入可能性の高い順に並べかえる.

２．予測上位$x$%中実際に購入した顧客を数える.

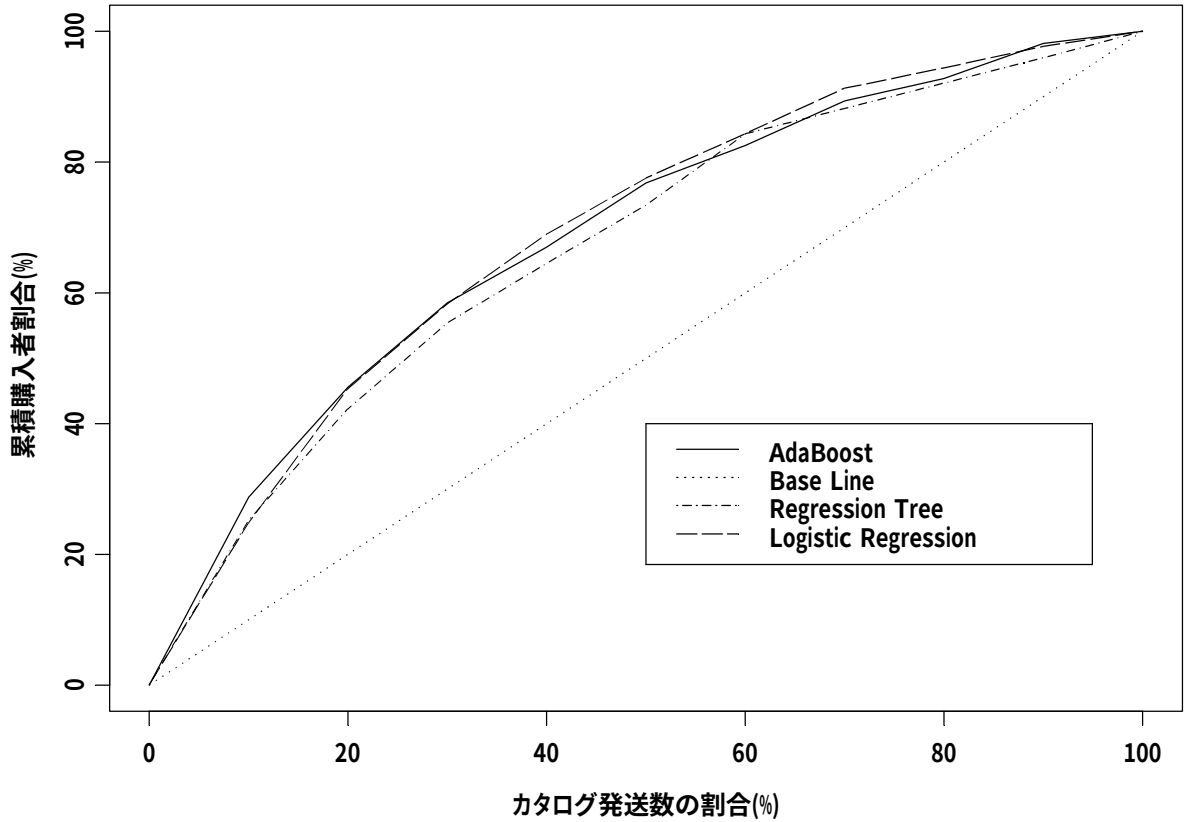
３．累積反応(購入)者割合を計算し横軸に顧客割合（$x$%）．縦軸に累積反応(購入)者割合をプロットする.

$$\text{累積反応者割合} = \frac{\text{予測上位 } x\%\text{中実際に購入した顧客の数}}{\text{全顧客中実際に購入した顧客の数}}$$

# テストデータにおけるゲインチャート（分析1）



# テストデータにおけるゲインチャート(分析2)



# テストデータにおける累積反応者割合（分析1）

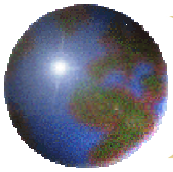
| | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| AdaBoost | **0.28** | **0.45** | 0.57 | 0.67 | 0.76 | 0.83 | 0.88 | **0.94** | **0.97** | 1.00 |
| 回帰木 | 0.24 | 0.43 | 0.56 | 0.67 | 0.72 | 0.77 | 0.83 | 0.89 | 0.95 | 1.00 |
| ロジスティック回帰 | 0.27 | 0.45 | **0.59** | **0.70** | **0.77** | **0.84** | **0.91** | 0.93 | 0.97 | 1.00 |

太字は最も各顧客割合において最も良かったスコアリングモデル

# テストデータにおける累積反応者割合（分析2）

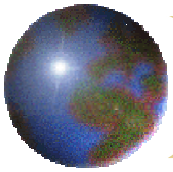
| | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| AdaBoost | **0.29** | **0.46** | **0.58** | 0.67 | 0.77 | 0.83 | 0.89 | 0.93 | **0.98** | 1.00 |
| 回帰木 | 0.25 | 0.42 | 0.55 | 0.65 | 0.73 | 0.84 | 0.88 | 0.92 | 0.96 | 1.00 |
| ロジスティック回帰 | 0.25 | 0.45 | 0.58 | **0.69** | **0.78** | **0.84** | **0.91** | **0.94** | 0.98 | 1.00 |

太字は最も各顧客割合において最も良かったスコアリングモデル



# ゲインチャートによる結論

- AdaBoostによる顧客スコアリングは上位2割の顧客の予測に優れており，実務上有効であると言える．
- 購入者割合を変更した分析（分析2）を行った結果，AdaBoostによるスコアリングモデルの性能が向上した．



# 利益分析 (1)

- スコアリングモデルから得た顧客ランキングをもとに何割の顧客を対象にマーケティングキャンペーンを行えばいいのか？
- 利益額の観点から最適なキャンペーンサイズを決め，その利益額をもとにスコアリングモデル間の比較を行う．

# 利益分析 (2)

**コンフュージョンマトリクス**

| | | 予測 | |
|---|---|---|---|
| | | 予測上位 $x$ % | 予測下位 $(1-x)$ % |
| 実際 | 購入 | $x_{11}$ | $x_{12}$ |
| | 非購入 | $x_{21}$ | $x_{22}$ |

**マーケティングキャンペーンの収支マトリックス**

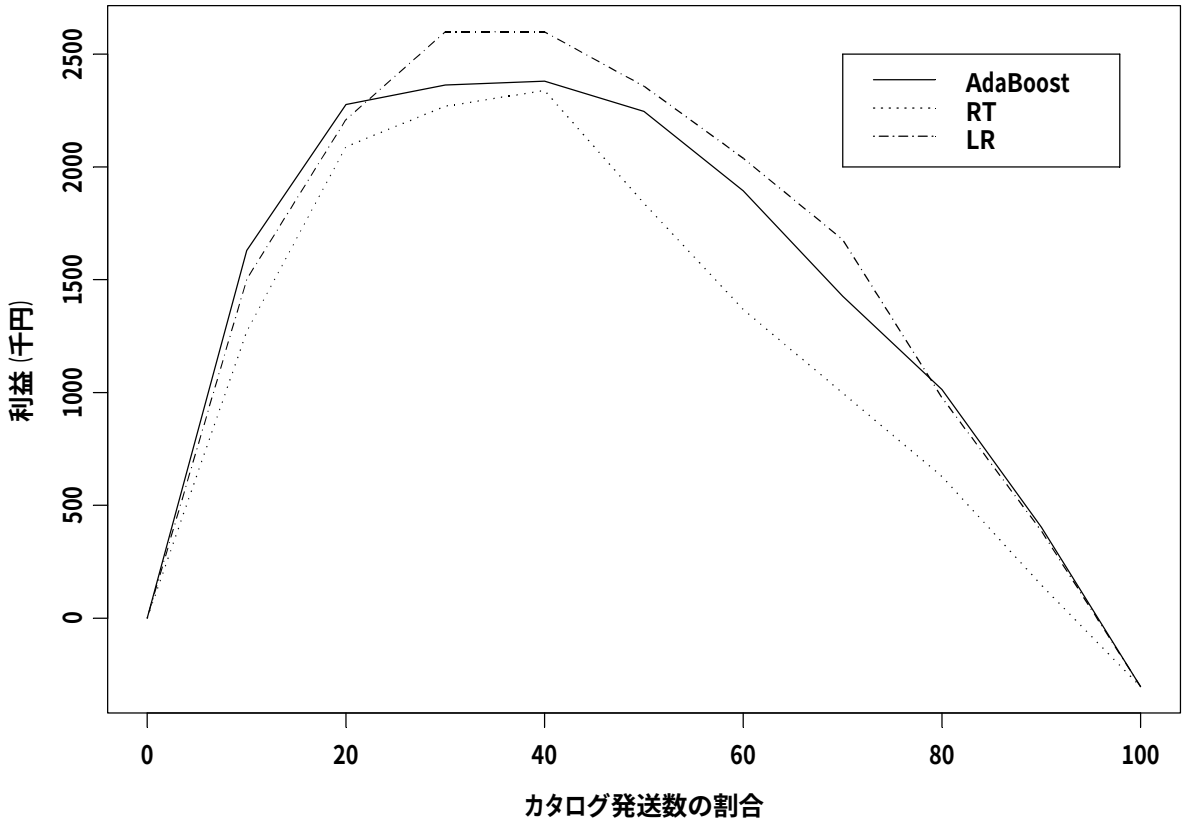
| | | 予測 | |
|---|---|---|---|
| | | 購入 | 非購入 |
| 実際 | 購入 | R-C (円) | 0 |
| | 非購入 | -C (円) | 0 |

**マーケティングキャンペーンの収益**

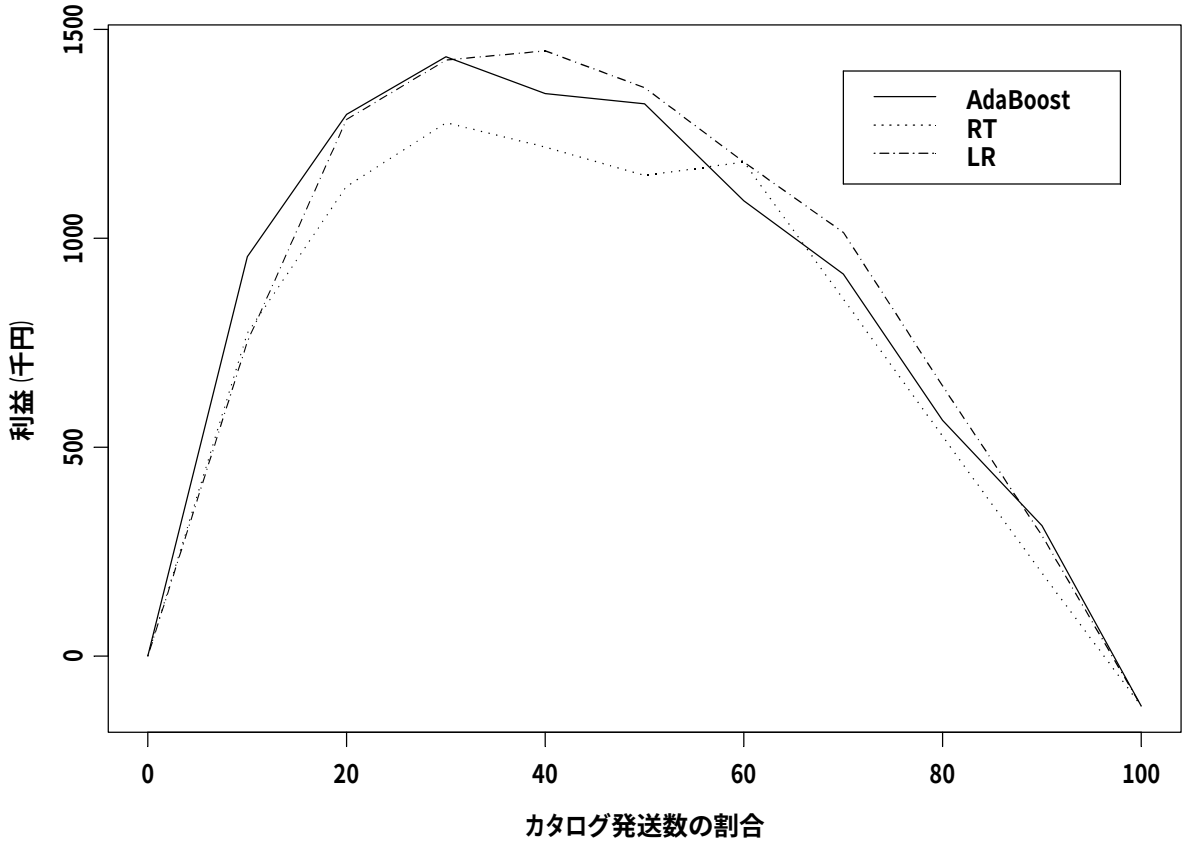
$$V = (R - C) \times x_{11} - C \times x_{21}$$

一人あたり平均収益R=10000, 一人あたり平均マーケティングコストC=200, 1000, 2000, 3000 の各場合の利益額を計算した.

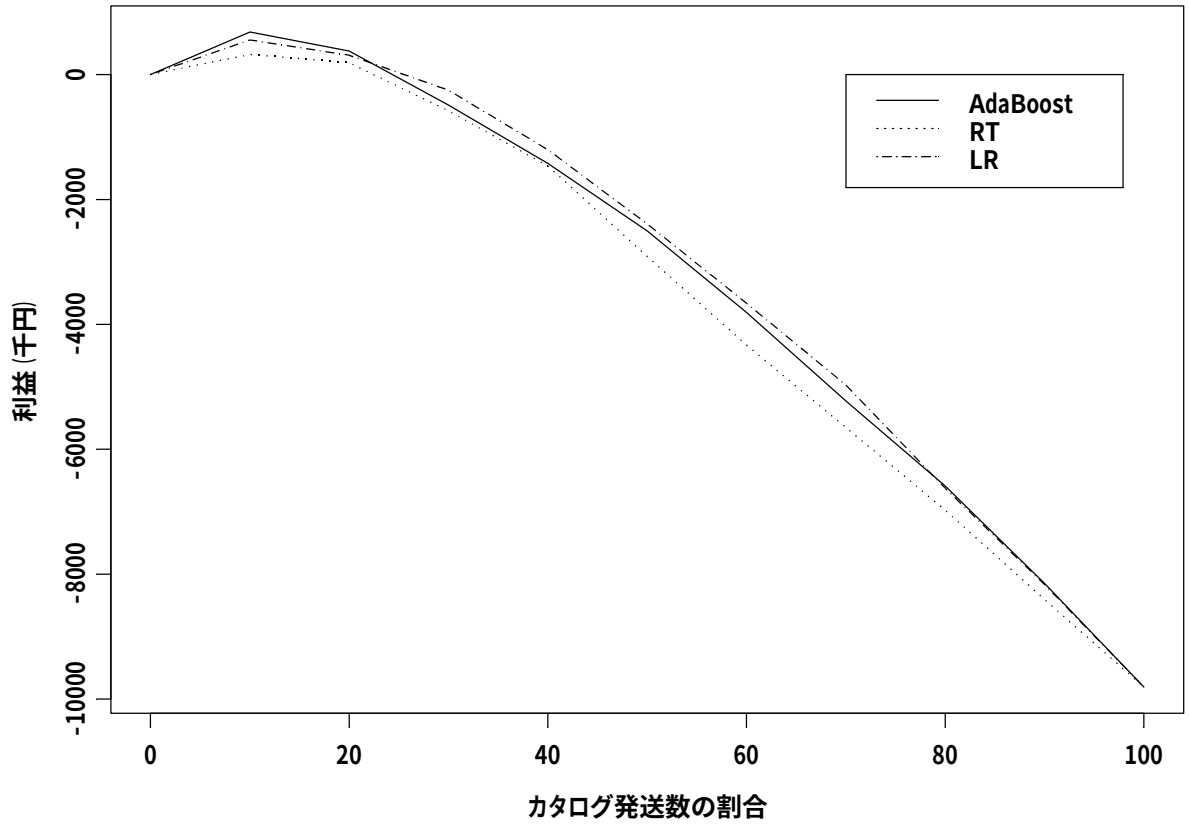
# C=1000の場合の利益分析（分析1）



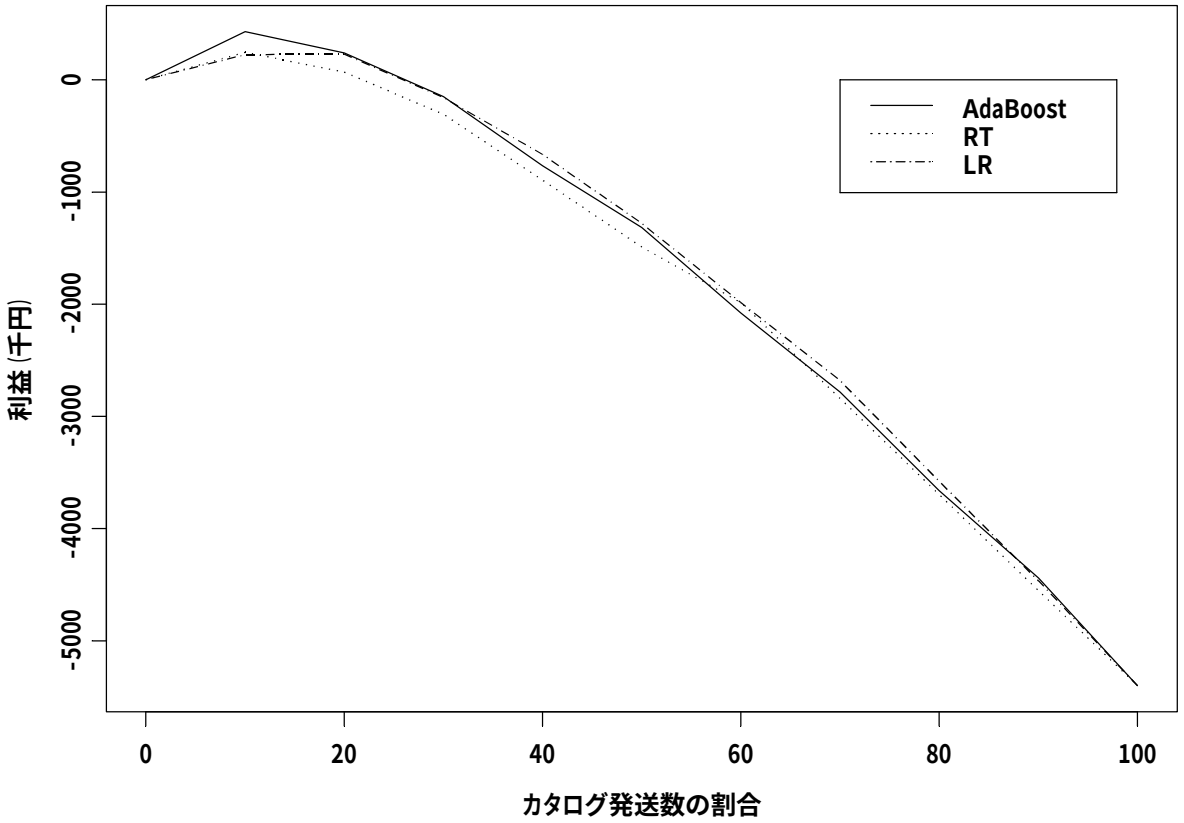
# C=1000の場合の利益分析（分析2）



# C=2000の場合の利益分析（分析1）



# C=2000の場合の利益分析（分析2）

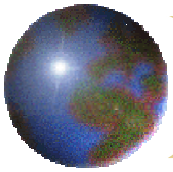


# 利益分析の結論

- 全てのスコアリングモデルにおいてC=200の場合，全ての顧客にカタログを送付し，C=3000の場合全ての顧客に送付しない方が利益額が高くなる結果となった．
- C=2000の場合，AdaBoostによるスコアリングモデルによって上位10%の顧客にカタログを送付すれば最も利益額が高くなることがわかった．

# 総合的な結論

- ブースティングによる顧客スコアリングは上位顧客の予測に有効であり，2：8の法則と合わせて考えると現実的に有効であると言える．
- 非購入者のデータが多い場合に無作為抽出によってデータ非購入者のデータの割合を減少させる分析手法は有効であることがわかった．
- 利益分析によってキャンペーンコストが高い場合にAdaBoostによるスコアリングは有効であることがわかった．



# 参考文献


- Freund, Y. and Schapire, R.E. “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, **55**, 119-139. 1997.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Additive logistic regression: a statistical view of boosting,” Technical Report, 1998.
- **竹林実，佐野夏樹，鈴木秀男**「AdaBoost**による顧客スコアリング**」，**日本オペレーションズ・リサーチ学会２００３年秋季研究発表会**
- Berry, M.J.A. and Linoff, G. **著，江原淳，金子武久，斉藤史郎，佐藤栄作，清水聰，寺田英治，守口剛共訳**「**マスタリング・データマイニングCRMのアートとサイエンス理論編**」（**海文堂**，2002）