Customizing Marketing Decisions Using Field Experiments

Spyros Zoumpoulis, INSEAD

Joint work with
Duncan Simester and Artem Timoshenko, MIT

Workshop on Data Driven Operations Management
EURANDOM, Eindhoven, The Netherlands
October 24, 2016

Motivation

Motivation

Motivation

Motivation

Motivation


COSTCO
WHOLESALE
WHY DOES COSTCO CHARGE
A MEMBERSHIP FEE?
Costco's membership fee provides a means of covering part of our operating costs and overheads thereby reducing our prices on the products we sell. This way the more members we have the lower our prices and the more you buy the more you save.

Motivation

Problem: customer segmentation

Motivation

Problem: customer segmentation

Questions

- Can you use data from a field experiment to target customers?
- How can you use data from a field experiment to train customer segmentation methods?

Data and Setup

- Large membership warehouse club chain
  - Annual membership: $50

Data and Setup

- Large membership warehouse club chain
  - Annual membership: \$50
- Two offers
  - \$25 membership (50% discount)
  - 120-day free trial membership (then need to pay full price)

Data and Setup

- Large membership warehouse club chain
  - Annual membership: $50
- Two offers
  - $25 membership (50% discount)
  - 120-day free trial membership (then need to pay full price)
- Stage 1 — Spring 2015
  - 1.2M households randomly assigned to 3 conditions: control, $25 paid offer, 120-day free trial.
  - 13 descriptive variables: housing characteristics, income and age characteristics, membership history, distances to closest retailer's and competitors' stores
  - Response variable: *profit* measure

Data and Setup

- Stage 2 — Fall 2015
    - Retailer chose to randomize in the *carrier-route* level

Data and Setup

- Stage 2 — Fall 2015
  - Retailer chose to randomize in the *carrier-route* level
  - 10,419 carrier routes corresponding to 4.1M households randomly assigned to 10 conditions:
    - 3 uniform policies (control, $25 discount, 120-day free trial)
    - 7 segmentation policies we propose
  - Same descriptive and response variables

Data and Setup

- Stage 2 — Fall 2015
  - Retailer chose to randomize in the *carrier-route* level
  - 10,419 carrier routes corresponding to 4.1M households randomly assigned to 10 conditions:
    - 3 uniform policies (control, \$25 discount, 120-day free trial)
    - 7 segmentation policies we propose
  - Same descriptive and response variables

- The learning problem
  - We aggregate Stage 1 data to 5,976 carrier routes: observations $\left(\mathbf{x}_i, y_i^{(control)}, y_i^{(\$25)}, y_i^{(120-\text{day})}\right)$
  - Use as training data for each of seven proposed segmentation methods
  - Apply segmentation methods to Stage 2 observations

Data and Setup

Stage 1 and Stage 2 datasets are different

Data and Setup

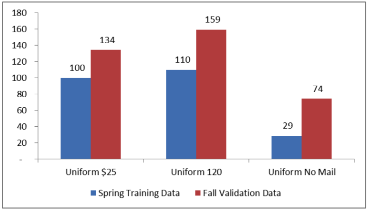
Stage 1 and Stage 2 datasets are different

	Stage 1	**Stage 2**
Time	spring 2015	fall 2015
Space	single geographic region	broader geographic area
Randomization	household level	carrier-route level
Mailing vehicle	covers of coupon book	postcard
Advertising	no campaign	mass media campaign

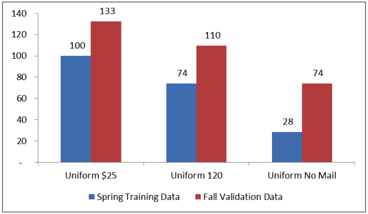
Data and Setup

Stage 1 and Stage 2 datasets are different

	Stage 1	**Stage 2**
Time	spring 2015	fall 2015
Space	single geographic region	broader geographic area
Randomization	household level	carrier-route level
Mailing vehicle	covers of coupon book	postcard
Advertising	no campaign	mass media campaign



Store revenue



Membership revenue

Roadmap

- Segmentation methods
  - Distance-driven policies
  - Model-driven policies
  - Classification policies
  - Uniform policies

Roadmap

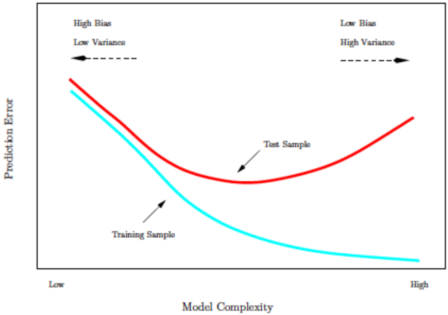
- Segmentation methods
    - Distance-driven policies
    - Model-driven policies
    - Classification policies
    - Uniform policies

Special care: Cross-validation

Roadmap

- Segmentation methods
  - Distance-driven policies
  - Model-driven policies
  - Classification policies
  - Uniform policies

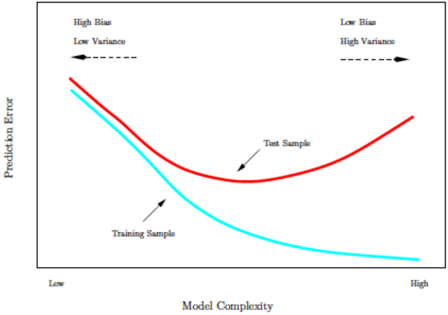
Special care: Cross-validation



Roadmap

- Segmentation methods
  - Distance-driven policies
  - Model-driven policies
  - Classification policies
  - Uniform policies

  Special care: Cross-validation



- Results

Kernel Regression

- Implementation
  Estimation of profit for Stage 2 observation $\mathbf{x}$ under treatment t:

$$\hat{y}^{(t)} = \frac{\sum_{i=1}^{N} K_\gamma(\mathbf{x}, \mathbf{x}_i) w_i^{(t)} y_i^{(t)}}{\sum_{i=1}^{N} K_\gamma(\mathbf{x}, \mathbf{x}_i) w_i^{(t)}},$$

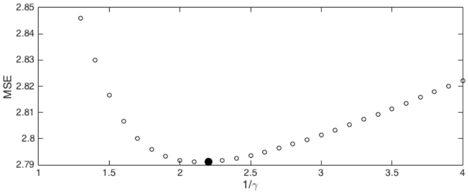
where $K_\gamma(\mathbf{x}, \mathbf{x}_i) = e^{-\gamma||\mathbf{x}-\mathbf{x}_i||^2}$

Kernel Regression

- Implementation
  Estimation of profit for Stage 2 observation $\mathbf{x}$ under treatment t:

$$\hat{y}^{(t)} = \frac{\sum_{i=1}^{N} K_\gamma(\mathbf{x}, \mathbf{x}_i) w_i^{(t)} y_i^{(t)}}{\sum_{i=1}^{N} K_\gamma(\mathbf{x}, \mathbf{x}_i) w_i^{(t)}},$$

  where $K_\gamma(\mathbf{x}, \mathbf{x}_i) = e^{-\gamma||\mathbf{x}-\mathbf{x}_i||^2}$
- Cross-validation

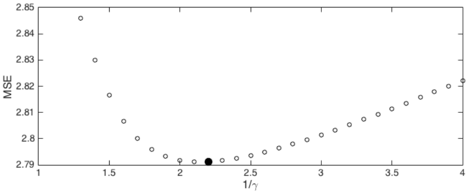


Kernel Regression

- Implementation
  Estimation of profit for Stage 2 observation $\mathbf{x}$ under treatment t:

$$\hat{y}^{(t)} = \frac{\sum_{i=1}^{N} K_\gamma(\mathbf{x}, \mathbf{x}_i) w_i^{(t)} y_i^{(t)}}{\sum_{i=1}^{N} K_\gamma(\mathbf{x}, \mathbf{x}_i) w_i^{(t)}},$$

  where $K_\gamma(\mathbf{x}, \mathbf{x}_i) = e^{-\gamma||\mathbf{x}-\mathbf{x}_i||^2}$
- Cross-validation



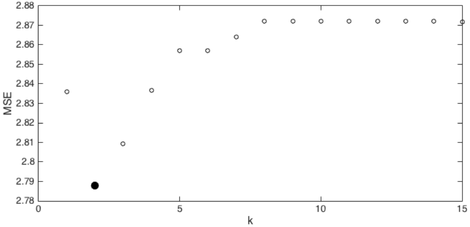
- Assignment
  For each new observation, predict profit for each treatment and choose best treatment

k-Nearest Neighbors

- Implementation
  Predict a Stage 2 observation's profit under each treatment by weight-averaging the profits of its *k nearest neighbors* from Stage 1

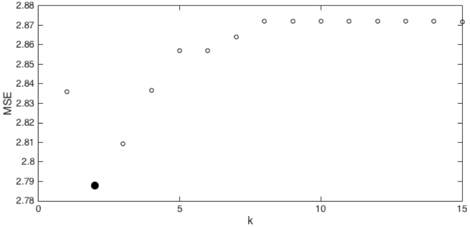
k-Nearest Neighbors

- Implementation
  Predict a Stage 2 observation's profit under each treatment by weight-averaging the profits of its *k nearest neighbors* from Stage 1
- Cross-validation



k-Nearest Neighbors

- Implementation
  Predict a Stage 2 observation's profit under each treatment by weight-averaging the profits of its *k nearest neighbors* from Stage 1
- Cross-validation



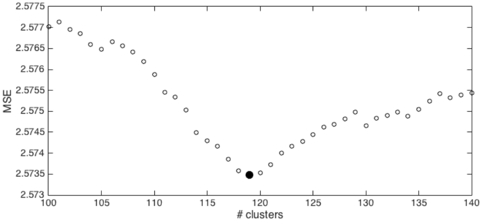
- Assignment
  For each new observation, predict profit for each treatment and choose best treatment

Hierarchical Clustering

- Implementation
  Cluster Stage 1 observations. Predict a Stage 2 observation's profit under each treatment by weight-averaging the profits of the Stage 1 observations in the *closest cluster*.

Hierarchical Clustering

- Implementation
  Cluster Stage 1 observations. Predict a Stage 2 observation's profit under each treatment by weight-averaging the profits of the Stage 1 observations in the *closest cluster*.
- Cross-validation

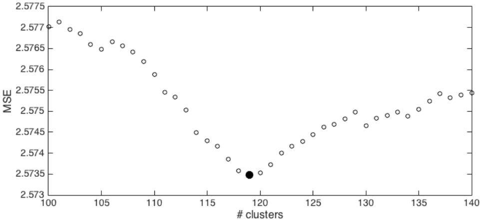


Hierarchical Clustering

- Implementation

  Cluster Stage 1 observations. Predict a Stage 2 observation's profit under each treatment by weight-averaging the profits of the Stage 1 observations in the *closest cluster*.

- Cross-validation



- Assignment

  For each new observation, predict profit for each treatment and choose best treatment

LASSO Regression

- Implementation

  The lasso regression estimates for treatment t are

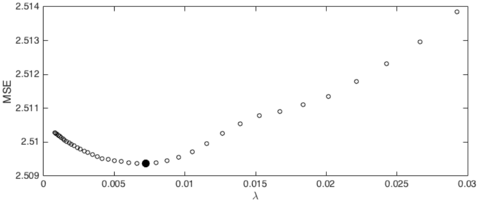
$$\beta^{(t)} = \arg\min_{\boldsymbol{\beta}^{(t)}} \left((\mathbf{y}^{(t)} - X\beta^{(t)})^T W^{(t)} (\mathbf{y}^{(t)} - X\beta^{(t)}) + \lambda ||\beta^{(t)}||_1 \right),$$

LASSO Regression

- Implementation
  The lasso regression estimates for treatment t are

$$\beta^{(t)} = \underset{\boldsymbol{\beta}^{(t)}}{\arg\min} \left((\mathbf{y}^{(t)} - X\beta^{(t)})^T W^{(t)} (\mathbf{y}^{(t)} - X\beta^{(t)}) + \lambda ||\beta^{(t)}||_1 \right),$$

- Cross-validation



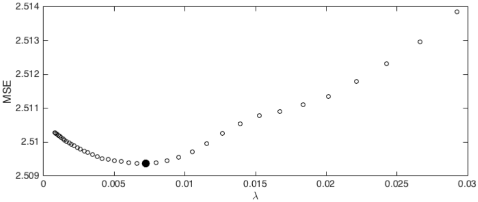
LASSO Regression

- Implementation

  The lasso regression estimates for treatment t are

$$\beta^{(t)} = \underset{\beta^{(t)}}{\arg\min} \left((\mathbf{y}^{(t)} - X\beta^{(t)})^T W^{(t)} (\mathbf{y}^{(t)} - X\beta^{(t)}) + \lambda ||\beta^{(t)}||_1 \right),$$

- Cross-validation



- Assignment

  For each new observation, predict profit for each treatment and choose best treatment

Finite Mixture Model

- Implementation
  Assume $y_i \sim f(y_i|\mathbf{x}_i;\boldsymbol{\theta};\boldsymbol{\pi}) = \sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell)$ with $\boldsymbol{\pi} \geq \mathbf{0}, \sum_{\ell=1}^{K} \pi_\ell = 1,$

Finite Mixture Model

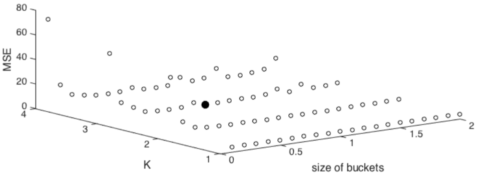
- Implementation
  Assume $y_i \sim f(y_i|\mathbf{x}_i;\boldsymbol{\theta};\boldsymbol{\pi}) = \sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell)$ with $\boldsymbol{\pi} \geq \mathbf{0}, \sum_{\ell=1}^{K} \pi_\ell = 1$,
  where $f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, $\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\theta}_\ell$.

Finite Mixture Model

- Implementation
  Assume $y_i \sim f(y_i|\mathbf{x}_i;\boldsymbol{\theta};\boldsymbol{\pi}) = \sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell)$ with $\boldsymbol{\pi} \geq \mathbf{0}, \sum_{\ell=1}^{K} \pi_\ell = 1$,
  where $f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, $\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\theta}_\ell$.
  Estimate with EM algorithm.

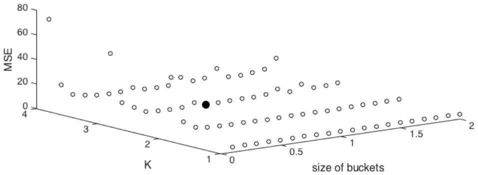
Finite Mixture Model

- Implementation
  Assume $y_i \sim f(y_i|\mathbf{x}_i;\boldsymbol{\theta};\boldsymbol{\pi}) = \sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell)$ with $\boldsymbol{\pi} \geq \mathbf{0}, \sum_{\ell=1}^{K} \pi_\ell = 1$,
  where $f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, $\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\theta}_\ell$.
  Estimate with EM algorithm.
- Cross-validation



Finite Mixture Model

- Implementation
  Assume $y_i \sim f(y_i|\mathbf{x}_i;\boldsymbol{\theta};\boldsymbol{\pi}) = \sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell)$ with $\boldsymbol{\pi} \geq \mathbf{0}, \sum_{\ell=1}^{K} \pi_\ell = 1$,
  where $f_\ell(y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, $\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\theta}_\ell$.
  Estimate with EM algorithm.
- Cross-validation



- Assignment
  For each new observation, predict profit for each treatment and choose best treatment

CHi-square Automatic Interaction Detection

- Implementation
  Decision tree for stage 1 observations: at each split, split the predictor that best explains the response variable if split, according to a chi-squared test for independence

CHi-square Automatic Interaction Detection

- Implementation
  Decision tree for stage 1 observations: at each split, split the predictor that best explains the response variable if split, according to a chi-squared test for independence
  - If independence, stop tree; else, create split and search for new split

CHi-square Automatic Interaction Detection

- Implementation
  Decision tree for stage 1 observations: at each split, split the predictor that best explains the response variable if split, according to a chi-squared test for independence
  - If independence, stop tree; else, create split and search for new split

- Cross-validation
  7 parameters: levels of significance for merging/splitting, number of observations in split s.t. no further split required, etc.

CHi-square Automatic Interaction Detection

- Implementation
  Decision tree for stage 1 observations: at each split, split the predictor that best explains the response variable if split, according to a chi-squared test for independence
  - If independence, stop tree; else, create split and search for new split
- Cross-validation
  7 parameters: levels of significance for merging/splitting, number of observations in split s.t. no further split required, etc.
- Assignment
  Assign each new observation the treatment that corresponds to the class the observation belongs to

Support Vector Machine

- Implementation

Support Vector Machine

- Implementation

  Label Stage 1 observations according to treatment with highest profit

Support Vector Machine

- Implementation

  Label Stage 1 observations according to treatment with highest profit

  Find maximally separating hyperplanes — 3-class classification

Support Vector Machine

- Implementation

Label Stage 1 observations according to treatment with highest profit

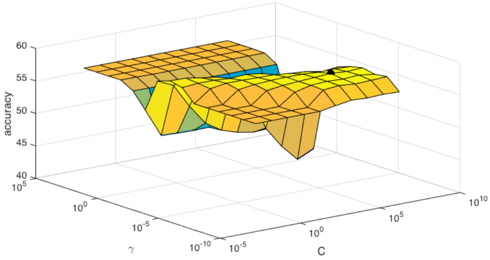
Find maximally separating hyperplanes — 3-class classification

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\theta_0,\boldsymbol{\xi}} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + C\sum_{i=1}^{N}\xi_i \\
\text{subject to} \quad & z_i(\boldsymbol{\theta}^T\boldsymbol{\phi}(\mathbf{x}_i) + \theta_0) \geq 1 - \xi_i, \\
& \xi_i \geq 0,
\end{aligned}
$$

where $K(\mathbf{x}_i, \mathbf{x}_i) = \boldsymbol{\phi}(\mathbf{x}_i)^T\boldsymbol{\phi}(\mathbf{x}_i)$ and $K_\gamma(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2}$

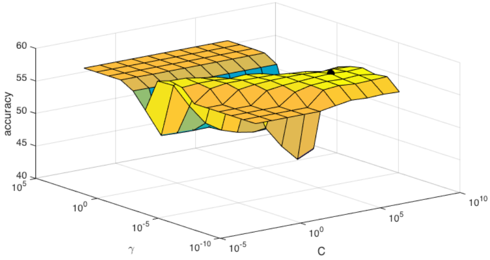
Support Vector Machine

- Cross-validation



Support Vector Machine

- Cross-validation



- Assignment

  Assign each new observation the treatment that corresponds to the class the observation belongs to

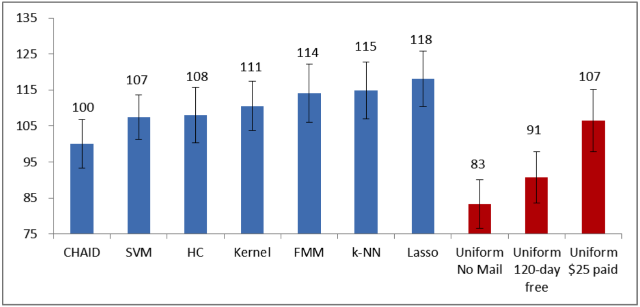
Uniform Policies

- Implementation and Assignment
  Each Stage 2 observation is assigned the same treatment

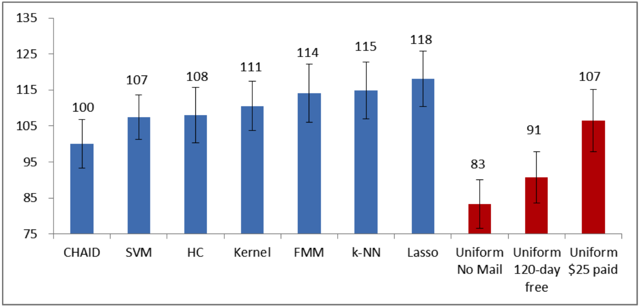
Uniform Policies

- Implementation and Assignment
  Each Stage 2 observation is assigned the same treatment
  - $25 paid 12-month membership policy
  - 120-day free trial membership policy
  - No-mail policy

Average Profit in Each Condition



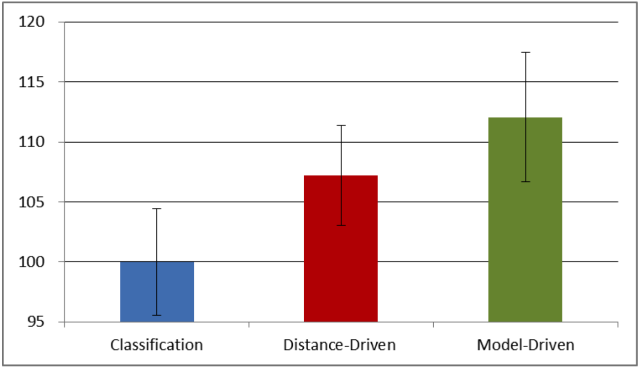
Average Profit in Each Condition



Lasso yields the highest average profit

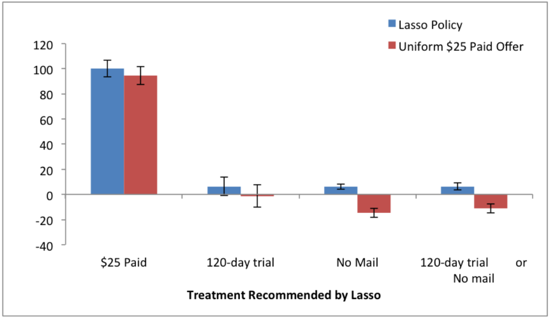
- Significantly higher than CHAID, SVM, Uniform policies ($p < 0.05$)

Average Profit in Each Condition



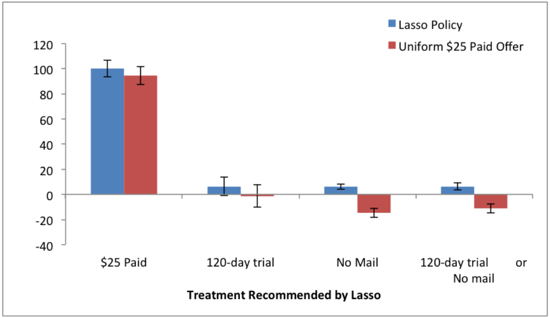
Model- and distance-driven methods significantly better than classifiers ($p < 0.01$)

Comparison with Uniform $25 Policy



- Where Lasso chose 120-day or no mail, it outperfomed the Uniform $25 policy significantly ($p < 0.01$)

Comparison with Uniform $25 Policy



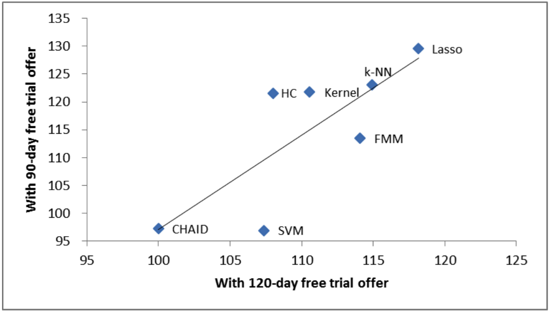
- Where Lasso chose 120-day or no mail, it outperfomed the Uniform $25 policy significantly ($p < 0.01$)
- Similarly for HC, Kernel, FMM, k-NN.
- CHAID and SVM: Uniform $25 policy better

Robustness

Stage 2 (fall): replace 120-day with 90-day free trial

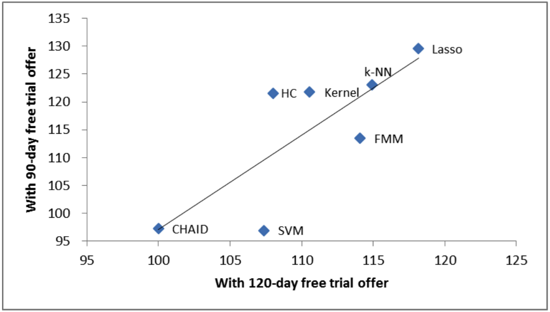
Robustness

Stage 2 (fall): replace 120-day with 90-day free trial



Robustness

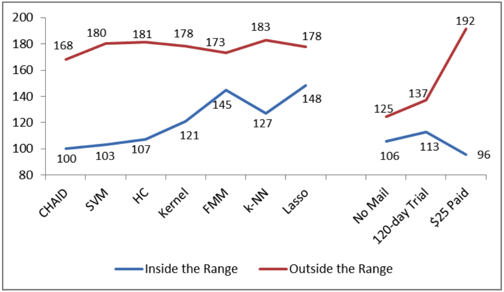
Stage 2 (fall): replace 120-day with 90-day free trial



Relative performance of methods robust to differences between the training data and the validation data

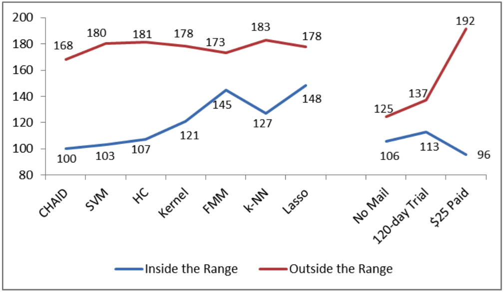
Performance Inside vs. Outside the Range of the Training Data

Stage 2 households *outside the range*: at least 1 of the 13 variables is at least 2 (Stage 1) st.dev.'s away from (Stage 1) mean



Performance Inside vs. Outside the Range of the Training Data

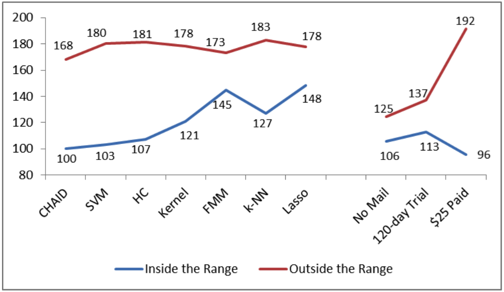
Stage 2 households *outside the range*: at least 1 of the 13 variables is at least 2 (Stage 1) st.dev.'s away from (Stage 1) mean



- Outside the range: all optimized methods perform similarly, worse than Uniform $25

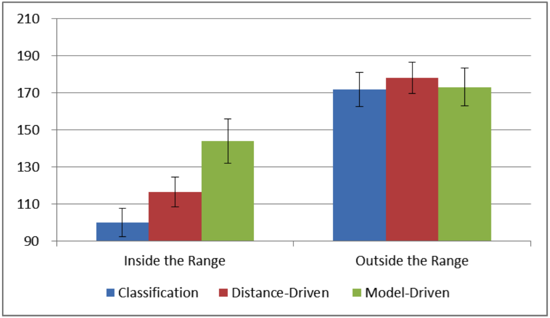
Performance Inside vs. Outside the Range of the Training Data

Stage 2 households *outside the range*: at least 1 of the 13 variables is at least 2 (Stage 1) st.dev.'s away from (Stage 1) mean



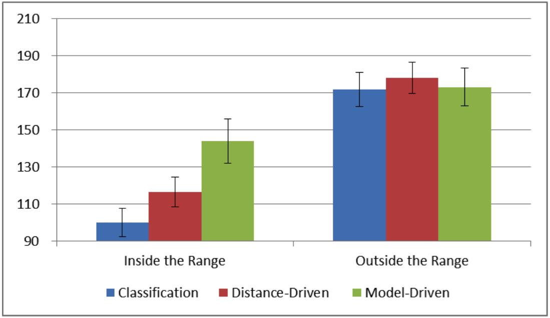
- Outside the range: all optimized methods perform similarly, worse than Uniform $25
- Inside the range: Lasso and FMM outperform other methods and uniform policies

Performance Inside vs. Outside the Range of the Training Data



- Outside the range: no significant differences

Performance Inside vs. Outside the Range of the Training Data



- Outside the range: no significant differences
- Inside the range: Model-driven > Distance-driven > Classifiers ($p < 0.01$)

Amount of Information: Size of Carrier Routes

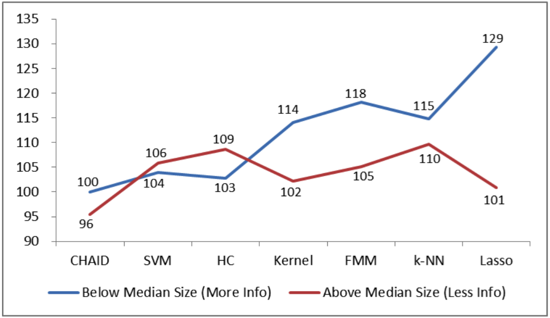
How does the precicion of information in the 13 variables affect outcomes?

- Stage 2 data: same set of 13 variables across households in a carrier route
- Amount of information for each household: varies with size of carrier route

Amount of Information: Size of Carrier Routes

How does the precicion of information in the 13 variables affect outcomes?

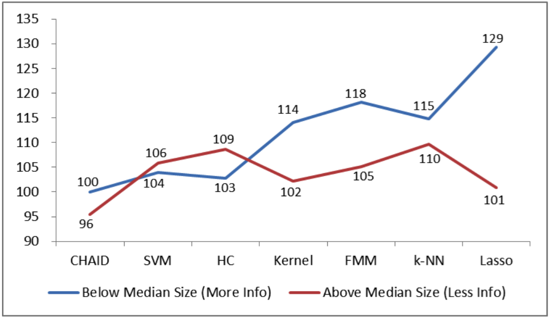
- Stage 2 data: same set of 13 variables across households in a carrier route
- Amount of information for each household: varies with size of carrier route



Amount of Information: Size of Carrier Routes

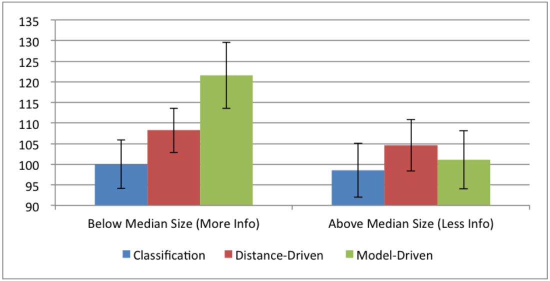
How does the precicion of information in the 13 variables affect outcomes?

- Stage 2 data: same set of 13 variables across households in a carrier route
- Amount of information for each household: varies with size of carrier route



- Above median size: all optimized methods perform similarly
- Below median size: Lasso performs significantly better

Amount of Information: Size of Carrier Routes



Model-based methods make the best use of the increased precision of information in smaller carrier routes

Discussion

- Model-driven methods perform the best
  - perform best in parts of the parameter space that are well represented in the training data
  - perform best when the information is more precise

Discussion

- Model-driven methods perform the best
  - perform best in parts of the parameter space that are well represented in the training data
  - perform best when the information is more precise
- Why do classifiers perform so badly?

Discussion

- Model-driven methods perform the best
  - perform best in parts of the parameter space that are well represented in the training data
  - perform best when the information is more precise

- Why do classifiers perform so badly?
  Loss of information: they only look *at which treatment* is optimal, and not *by how much*

Discussion

- Model-driven methods perform the best
  - perform best in parts of the parameter space that are well represented in the training data
  - perform best when the information is more precise

- Why do classifiers perform so badly?
  Loss of information: they only look *at which treatment* is optimal, and not *by how much*
  Example: mailing costs $1. No response w.p. 0.95, profit of 1000 w.p. 0.05

Discussion

- Model-driven methods perform the best
  - perform best in parts of the parameter space that are well represented in the training data
  - perform best when the information is more precise

- Why do classifiers perform so badly?
  Loss of information: they only look *at which treatment* is optimal, and not *by how much*
  Example: mailing costs \$1. No response w.p. 0.95, profit of 1000 w.p. 0.05
  - Regression-based methods: Choose treatment with higher expected profit and mail to all

Discussion

- Model-driven methods perform the best
  - perform best in parts of the parameter space that are well represented in the training data
  - perform best when the information is more precise

- Why do classifiers perform so badly?
  Loss of information: they only look *at which treatment* is optimal, and not *by how much*
  Example: mailing costs \$1. No response w.p. 0.95, profit of 1000 w.p. 0.05
  - Regression-based methods: Choose treatment with higher expected profit and mail to all
  - Classifiers: Choose treatment that is optimal more frequently, and never mail

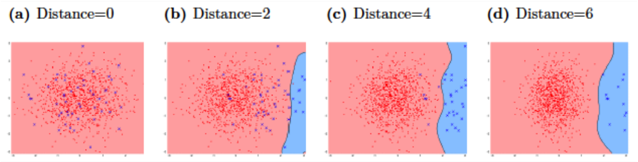
Discussion

- Model-driven methods perform the best
  - perform best in parts of the parameter space that are well represented in the training data
  - perform best when the information is more precise

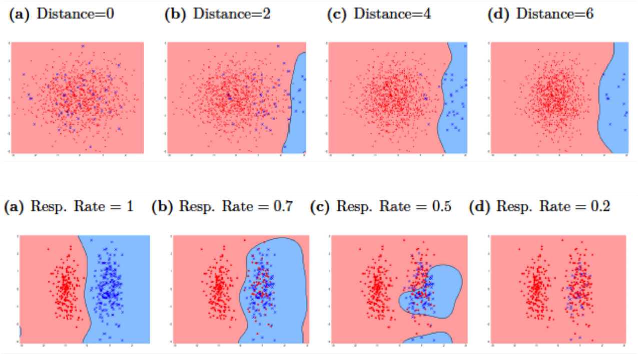
- Why do classifiers perform so badly?
  Loss of information: they only look *at which treatment* is optimal, and not *by how much*
  Example: mailing costs $1. No response w.p. 0.95, profit of 1000 w.p. 0.05
  - Regression-based methods: Choose treatment with higher expected profit and mail to all
  - Classifiers: Choose treatment that is optimal more frequently, and never mail

- But classifiers do well when
  - descriptive variables distinguish the segments
  - outperformance margins between treatments are symmetric.

Discussion

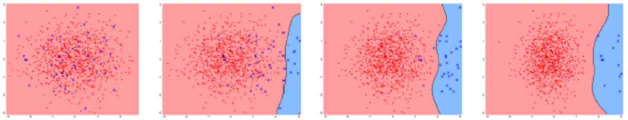


Discussion

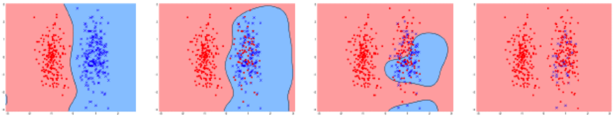

(a) Distance=0
(b) Distance=2
(c) Distance=4
(d) Distance=6
(a) Resp. Rate = 1
(b) Resp. Rate = 0.7
(c) Resp. Rate = 0.5
(d) Resp. Rate = 0.2

Discussion

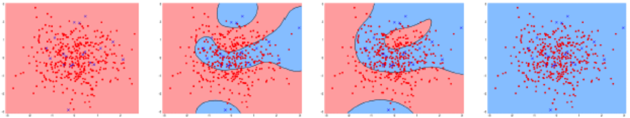
(a) Distance=0 **(b)** Distance=2 **(c)** Distance=4 **(d)** Distance=6



(a) Resp. Rate = 1 **(b)** Resp. Rate = 0.7 **(c)** Resp. Rate = 0.5 **(d)** Resp. Rate = 0.2



(a) Cost =[1,1] **(b)** Cost =[1,20] **(c)** Cost =[1,100] **(d)** Cost =[1,1000]



The Value of Field Experiments

- Can we make marketing decisions with a practically feasible number of field experiments?

  Yes. (Management Science, *The value of field experiments*)

The Value of Field Experiments

- Can we make marketing decisions with a practically feasible number of field experiments?
  Yes. (Management Science, *The value of field experiments*)

- Can we use data from field experiments to target customers?
  Yes. This paper

The Value of Field Experiments

- Can we make marketing decisions with a practically feasible number of field experiments?
  Yes. (Management Science, *The value of field experiments*)

- Can we use data from field experiments to target customers?
  Yes. This paper

- Can we use field experiments to optimize a sequence of promotions/to retarget non-respondents?
  Yes. Current work