Applications of regression, classification and likelihood maximization

IDS 575 – March 18, 2019

Prof. Theja Tulabandhula

Guest lecture by: Parshan Pakiman (ppakim2@uic.edu, homepage)

Agenda

- **Introduction**: brief review of regression, classification, and likelihood maximization (6 slides)
- **Application 1 (OM/OR):** newsvendor problem (12 slides)
- **Application 2 (Healthcare):** a mortality prediction problem (9 slides)
- **Break**: 10 minutes break
- **Application 3 (Marketing):** shopper marketing optimization (20 slides)

Some Remarks

1. For slides corresponding to the shopper marketing application, please contact Parshan Pakiman at *ppakim2@uic.edu*.

2. There is a feedback form that you can fill after the lecture. (Here is a link to the form. The link is also shared on the IDS-575 forum)

3. Filling the form is not required, but definitely your feedback is important and helpful for me.

4. For all assignments, make sure that you turn in both **PDF/HTML** and **Rmd** files.

5. Please answer questions on final exam **briefly** and **concisely**.

6. Feel free to bring up questions to have discussion!

7. Please stop me if you have any question!

Introduction

Introduction

UIC BUSINESS

1. Regression, classification, and MLE are methods for estimating parameters of predictive models.
2. They have been applied to various problems in different fields such as
    - Accounting and Finance
    - Marketing
    - Operations Management and Operations Research
    - Biological Science and Biomedical Engineering
    - Psychological Science
    - Healthcare
    - …

Let's briefly review these models!

Some Notations

UIC BUSINESS

- Predictors, independent variable, features, and covariates:

$$X = (X_1, X_2, \ldots, X_p) \quad p\text{-dimensional vector}$$

- Target variable, dependent variable, and label:

$$Y \text{ or } f(X)$$

- A model:

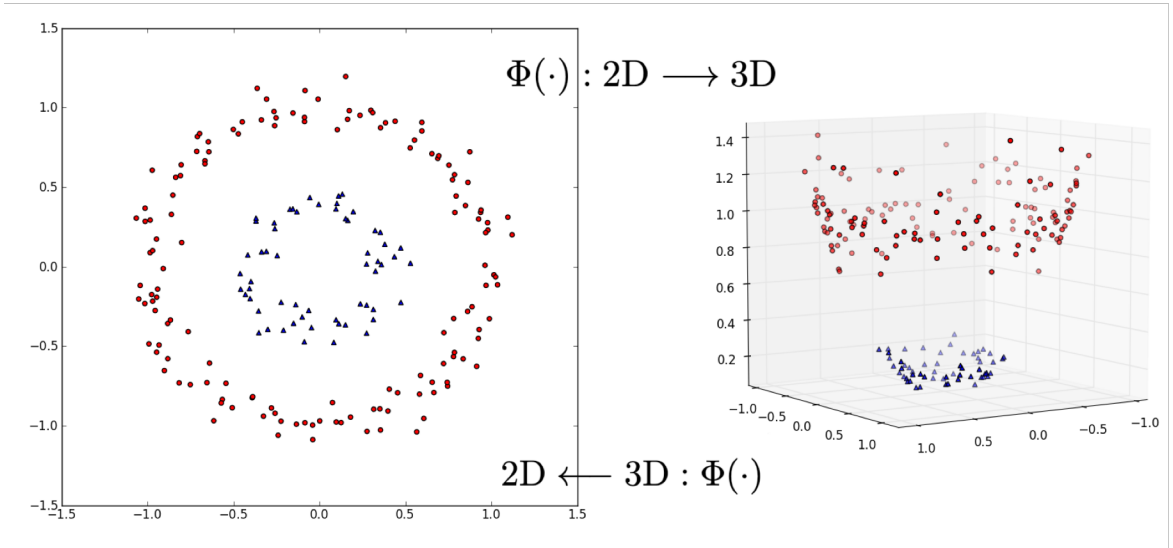
$$Y = f(X)$$

for example,

$$f(X) = \beta^\top X$$

$$f(X) = \beta_0 + \beta^\top X$$

$$f(X) = \beta_0 + \beta^\top \Phi(X)$$

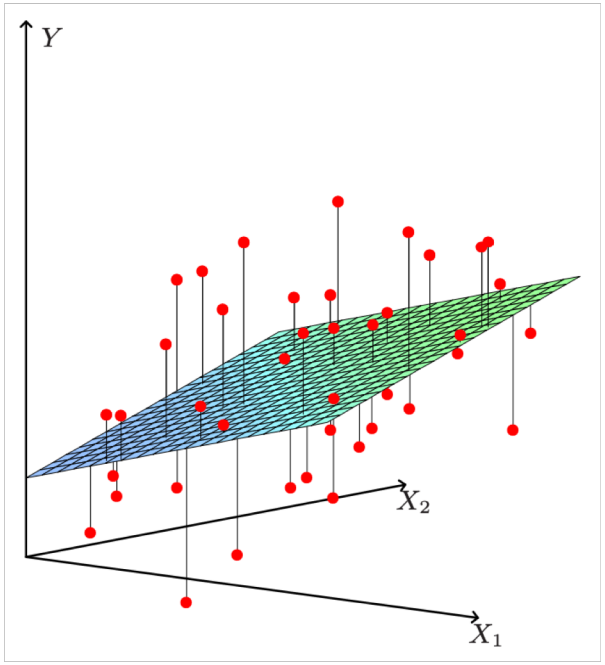


Regression

UIC BUSINESS

Predictors:

$$X = (X_1, X_2, \ldots, X_p)$$



Linear regression model:

$$f(X) = \beta_0 + \beta^\top X = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Unknows!

$$\hat{X} = \begin{bmatrix} \hat{X}_{1,1} & \hat{X}_{1,2} & \cdots & \hat{X}_{1,p} \\ \hat{X}_{2,1} & \hat{X}_{2,2} & \cdots & \hat{X}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{X}_{n,1} & \hat{X}_{n,2} & \cdots & \hat{X}_{n,p} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad f(\hat{X}) = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}.$$

Classification

UIC BUSINESS

Example[1]:

Dataset: 130, 000 images of skin lesions/2, 000 different diseases

- Test data: 370 high-quality, biopsy-confirmed images
- Better performance than 23 Stanford dermatologists
- 10,000 hours no match for deep learning and large datasets

Some classifiers:

- Classification with linear regression
- Gradient boosting
- Logistic classifier
- Naïve Bayes classifier
- Linear discriminant analysis
- Support vector machines

1. http://faculty.chicagobooth.edu/nicholas.polson/teaching/41900/beamer41901-2018.pdf

MLE/MAE

Family of parameterized distributions:

$$\left\{ f(\cdot, \theta) \ : \ f \text{ is a joint distribution} \right\}$$

$$f(X, \theta = (\mu, \sigma)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}, \qquad X \text{ is a 1D vector}$$

Likelihood function: $X \mapsto f(X \mid \theta)$

MLE: find a θ that maximizes $\max_{\theta} \log f(X|\theta)$

MAE: find a θ that maximizes $\max_{\theta} \log f(X|\theta)p(\theta)$

Prior knowledge

Newsvendor Problem

Newsvendor (NV) Problem

UIC BUSINESS

Problem statement[1]: The Fashion Store sells fashion items. The store has to order these items many months in advance of the fashion season in order to get a good price on the items. Each unit costs Fashion $100. These units are sold to customers at a price of $250 per unit. Items not sold during the season can be sold to the outlet store at $80 per unit. If the store runs out of an item during the season it has to obtain the item from alternative sources and the cost including air freight to Fashion is $190 per unit.

Fashion wants help in choosing the initial order quantity to maximize net contribution from running the store. What should we do?

1. The example is adopted from 'http://faculty.chicagobooth.edu/donald.eisenstein/research/NewsvendorModel.pdf'

Newsvendor Problem

UIC BUSINESS

Let's do some calculation:

Excess cost: $c_e = 100 - 80 = 20$

Shortage cost: $c_s = (250 - 100) - (250 - 190) = 90$

Unit costs = \$100
Price for customers = \$250
Price in outlet = \$80
Replenishment cost = \$190

What is expected cost of ordering z units?

Expected marginal cost = $\mathrm{P}(\text{Demand} \leq z)c_e = 20 \times \mathrm{P}(\text{Demand} \leq z)$

Expected marginal profit = $\mathrm{P}(\text{Demand} \geq z)c_s = 90 \times \mathrm{P}(\text{Demand} \leq z)$

Case 1. expected marginal cost > expected marginal profit?

Decrease order level z

Case 2. expected marginal cost < expected marginal profit?

Increase order level z

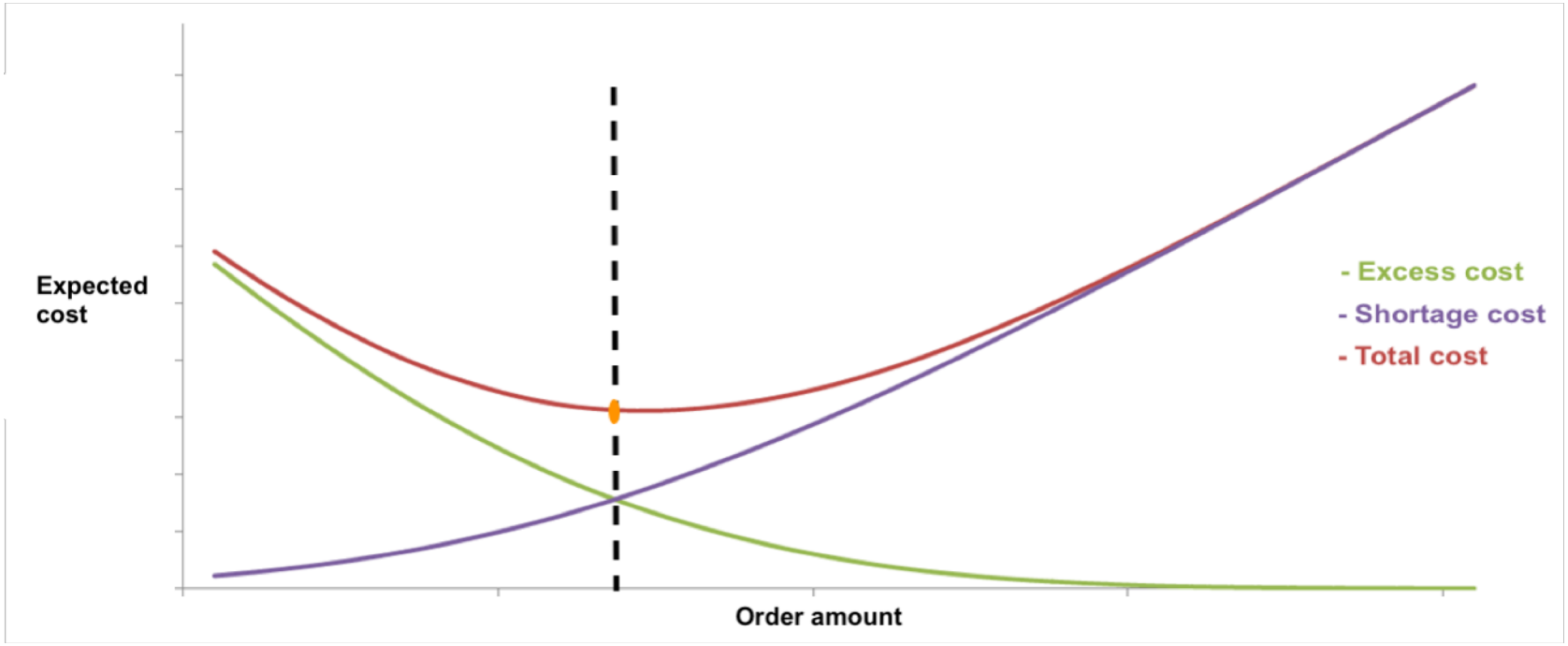
Optimal Order Level

UIC BUSINESS

Case 3. expected marginal cost = expected marginal profit

Optimal order level z^*

Increase → | ← Decrease



Optimal ordering quantity

Case 3. expected marginal cost = expected marginal profit

$$\mathrm{P}(\text{Demand} \geq z^*)c_s = \mathrm{P}(\text{Demand} \leq z^*)c_e$$

$$(1 - \mathrm{P}(\text{Demand} \leq z^*))c_s = \mathrm{P}(\text{Demand} \leq z^*)c_e$$

$$\mathrm{P}(\text{Demand} \leq z^*) = \frac{c_s}{c_s + c_e}$$

Optimal solution:

$$z^* = \inf\left\{ z : \mathrm{P}(\text{Demand} = z) \geq \frac{c_s}{c_s + c_e}) \right\}$$

NV Math Programs

UIC BUSINESS

- (NV) The NV problem can thus be written as $\min_z \ \mathbb{E}[c(z, D)]$

  where the cost function is $c(z, D) = c_s(D - z)^+ + c_e(z - D)^+$

$$z^* = \inf \left\{ z : F_{\text{Demand}}(z) \geq \frac{c_s}{c_s + c_e}) \right\}$$

- Data-driven NV) For a given set of realized demands $\{d_1, \ldots, d_N\}$

  we need to solve $\min_z \frac{1}{N} \sum_{i=1}^{N} \left(c_s(d_i - z)^+ + c_e(z - d_i)^+ \right)$

$$z^* = \inf \left\{ z : \widehat{F}_{\text{Demand}}(z) \geq \frac{c_s}{c_s + c_e}) \right\}$$

1. What if we have some predictors for demand?
2. How can we incorporate them into the NV model?

The Big Data Newsvendor[1] – I

UIC BUSINESS

Assume we have dataset $S_n = \{(d_i, X_i) : i = 1, 2, \ldots, N\}$ where X_i is a p-dimensional feature vector belonging to set $\mathcal{X}$

(NV-features) We can incorporate predictors into the data-driven NV model follows,

$$\min_{z(\cdot):\mathcal{X}\to\mathbb{R}} \frac{1}{N}\sum_{i=1}^{N} c_s(d_i - z(X_i))^+ + c_e(z(X_i) - d_i)^+$$

We can use idea of regression and write $z(X) = \beta^\top X$ or $\beta^\top \Phi(X)$

$$\min_{\beta} \frac{1}{N}\sum_{i=1}^{N} c_s(d_i - \beta^\top X_i)^+ + c_e(\beta^\top X_i - d_i)^+$$

The above math program can be solved via gradient descent or linear programming

1. Rudin, Cynthia, and Gah-Yi Vahn. "The big data newsvendor: Practical insights from machine learning." (2014).

The Big Data Newsvendor[1] – II

UIC BUSINESS

NV-LP: following linear programming computes coefficients in the linear model:

$$\min_{\beta_1,\ldots,\beta_p} \frac{1}{N}\sum_{i=1}^{N} c_s s_i + c_e t_i$$

$$s_i \geq d_i - \beta^\top X_i, \qquad \forall i = 1, 2, \ldots, N$$

$$t_i \leq \beta^\top X_i - d_i, \qquad \forall i = 1, 2, \ldots, N$$

$$s_i, t_i \geq 0$$

- The above problem can be solved efficiently.
- We can also use LASSO regularizer and compute coefficients via linear programming.
- For Ridge regularizer, we need to solve a quadratic program with linear constraints.

1. Rudin, Cynthia, and Gah-Yi Vahn. "The big data newsvendor: Practical insights from machine learning." (2014).

Closed Form Solutions Summary

UIC BUSINESS

- Knowing true demand distribution:

$$z^* = \inf \left\{ z : F_{\text{Demand}}(z) \geq \frac{c_s}{c_s + c_e}) \right\}$$

- Empirical demand distribution:

$$z^* = \inf \left\{ z : \widehat{F}_{\text{Demand}}(z) \geq \frac{c_s}{c_s + c_e}) \right\}$$

- Empirical demand distribution with binary features:

$$z_0^* = \inf \left\{ z : \widehat{F}_{\text{Demand}|X=0}(z) \geq \frac{c_s}{c_s + c_e} \right\}$$

$$z_0^* + z_1^* = \inf \left\{ z : \widehat{F}_{\text{Demand}|X=1}(z) \geq \frac{c_s}{c_s + c_e} \right\}$$

Numerical Study

UIC BUSINESS

No. of past days	without OS Features		with OS Features	
	Median Cost as % of SAA (p-value)	Total no. of Features (avg. chosen)	Median Cost as % of SAA (p-value)	Total no. of Features (avg. chosen)
0	61.90 (0.000)	12 (3.00)	61.90 (0.000)	12 (3.00)
1	63.07 (0.000)	15 (4.00)	62.73 (0.000)	27 (6.28)
2	61.90 (0.000)	27 (5.31)	67.17 (0.000)	51 (9.11)
3	61.84 (0.000)	39 (6.23)	57.24 (0.000)	75 (21.12)
4	57.53 (0.000)	51 (7.97)	57.45 (0.000)	99 (28.07)
5	**57.50 (0.000)**	**63 (8.25)**	56.63 (0.000)	123 (36.68)
6	58.29 (0.000)	75 (8.63)	58.79 (0.000)	147 (43.36)
7	57.85 (0.000)	87 (9.53)	58.53 (0.000)	171 (52.41)
8	57.63 (0.000)	99 (9.52)	**54.31 (0.000)**	**195 (56.51)**
9	57.64 (0.000)	111 (9.56)	58.19 (0.000)	219 (62.62)
10	57.70 (0.000)	123 (9.55)	61.23 (0.000)	243 (69.68)
11	57.64 (0.000)	135 (9.68)	60.05 (0.000)	267 (73.43)
12	57.64 (0.000)	147 (9.85)	66.13 (0.000)	291 (82.24)
13	57.88 (0.000)	159 (10.33)	59.22 (0.000)	315 (91.78)
14	57.67 (0.000)	171 (10.30)	60.54 (0.000)	339 (97.22)

Table 2 The median out-of-sample cost of (NV-algo) relative to SAA on the validation dataset. We use day of the week, time of the day and x number of days of past demand as features. The column marked "with OS" refer to results using differences of past demands as features, as inspired by OS. The best results without and with OS are highlighted in bold. In parentheses we report the p-values from the Wilcoxon rank-sum test to compare the result against SAA.

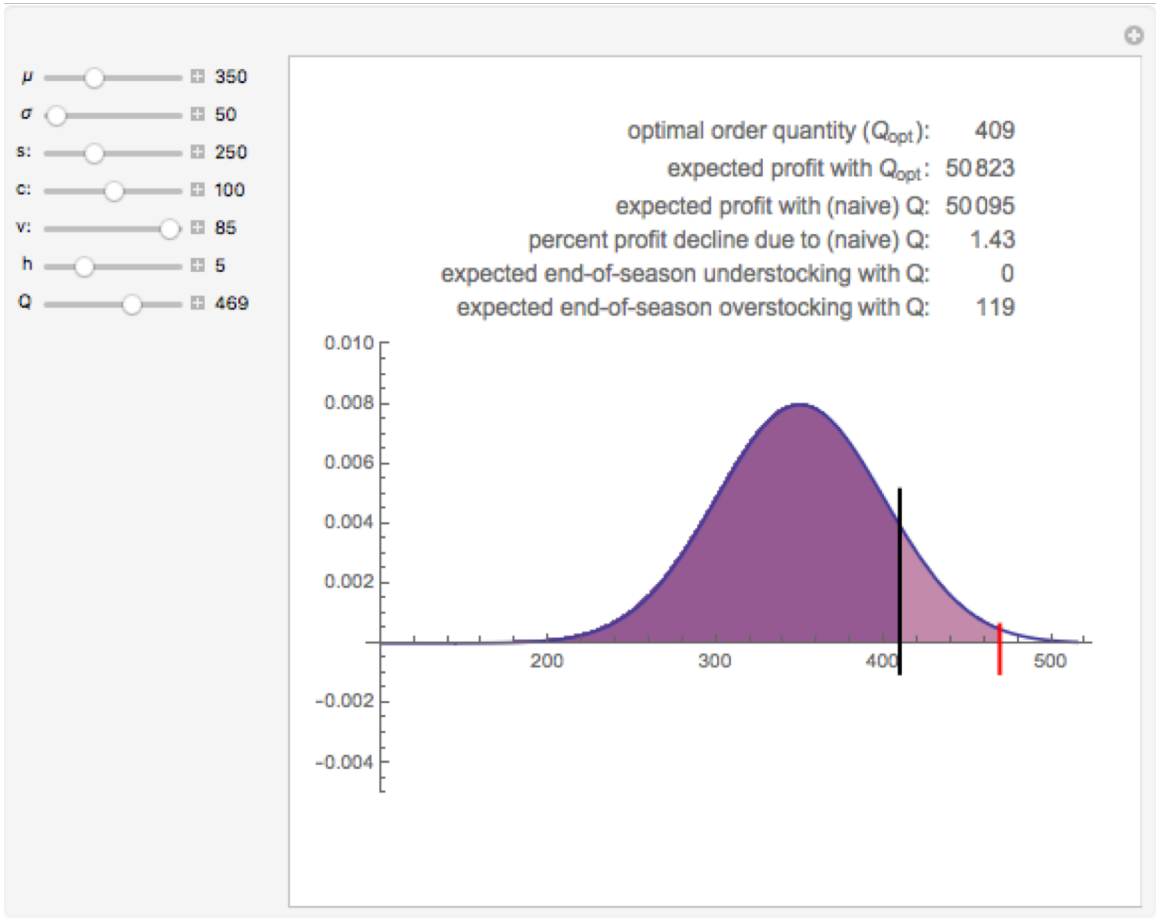
1.

Table is copied from Rudin, Cynthia, and Gah-Yi Vahn. "The big data newsvendor: Practical insights from machine learning." (2014).

Newsvendor Simulator[1]

UIC BUSINESS

Playing with NV simulator:



μ: average (mean) demand

σ: standard deviation of demand

s: unit selling price

c: unit purchase cost

v: unit markdown price for unsold items

h: unit holding cost for unsold items

Q: naive (or heuristic) order quantity

1. https://www.wolframcloud.com/objects/demonstrations/CapacityPlanningForShortLifeCycleProductsTheNewsvendorModel-source.nb

Mortality Prediction

Mortality Prediction of ICU Patients[1]

UIC BUSINESS

- Quantifying patient health and predicting future outcomes is critical.
- There are many works for constructing predictive model for patient health condition in ICU.
- Medical Information Mart for Intensive Care (MIMIC) database has a huge information about patients in ICU.
- It has 40 tables, 50K patient admissions, and about 730M records.
- Feature engineering is important for this huge database for obtaining good out-of-sample performance.
- Different methods such as logistic regression and gradient boosting have been used.

1. Johnson, Alistair EW, Tom J. Pollard, and Roger G. Mark. "Reproducibility in critical care: a mortality prediction case study." *Machine Learning for Healthcare Conference*. 2017.

- **Time window:** different cohort time windows can be considered such as the baseline cohort window that begins at ICU admission and ends up to 24 hours after ICU admission
- **Vital sign measurements:** heart rate, blood pressure, respiratory rate, oxygen saturation. Features like *first, last, minimum, and maximum* value across the window can be created.
- **laboratory measurements:** since laboratory measurements for patients are usually available before entering ICU, time window extended the window backwards by 24 hours and extracted the first and last measurement

Some Created Features

UIC BUSINESS

Time window	Feature extracted	Variables
$[t_{i,w} - W, t_{i,w}]$	Minimum, Maximum, First, Last	Heart rate, Systolic/Diastolic/Mean blood pressure, Respiratory rate, Temperature, Peripheral Oxygen Saturation, Glucose
$[t_{i,w} - W, t_{i,w}]$	Minimum	Glasgow coma scale
$[t_{i,w} - W, t_{i,w}]$	Last	Glasgow coma scale, Glasgow coma scale components (motor, verbal, eyes), unable to collect verbal score
$[t_{i,w} - W - 24, t_{i,w}]$	First, last	Oxygen saturation, Partial pressure of oxygen, Partial pressure of carbon dioxide, Arterial-alveolar gradient, Ratio of partial pressure of oxygen to fraction of oxygen inspired, pH, Base excess, Bicarbonate, Total carbon dioxide concentration, Hematocrit, Hemoglobin, Carboxyhemoglobin, Methemoglobin, Chloride, Calcium, Temperature, Potassium, Sodium, Lactate, Glucose
$[t_{i,w} - W - 24, t_{i,w}]$	First, last	Anion gap, Albumin, Immature band forms, Bicarbonate, Bilirubin, Creatinine, Chloride, Glucose, Hematocrit, Hemoglobin, Lactate, Platelet, Potassium, Partial thromboplastin time, International Normalized Ratio, Prothrombin time, Sodium, Blood urea nitrogen, White blood cell count
$[t_{i,w} - W - 24, t_{i,w}]$	Sum	Urine output

Models and Evaluation

UIC BUSINESS

- **Target variable:** in-hospital mortality, 30-day post ICU admission mortality, 48-hour post ICU discharge mortality, 30-day post ICU discharge mortality, 30-day post hospital discharge mortality, 6-month post hospital discharge mortality, 1-year post hospital discharge mortality, and 2 year post hospital discharge mortality
- **Models**: gradient boosting and logistic regression have been implemented in python (code is available on github[1]).
- **Evaluation:** 5-fold cross-validation to obtain estimates of model performance and AUC has been reported.

1. https://github.com/alistairewj/mortality-prediction

Let's See a Video …

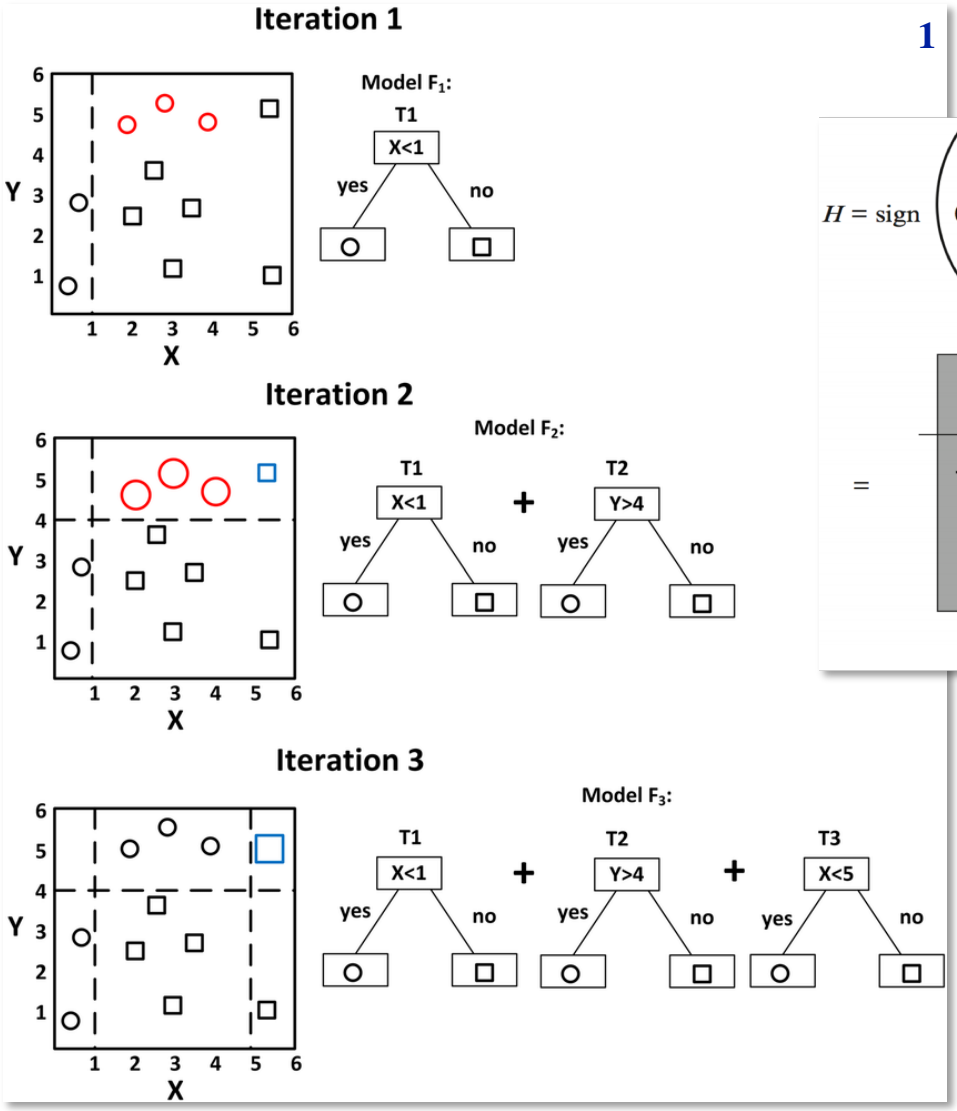
UIC BUSINESS

Ghassemi, Marzyeh, et al. "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data." *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

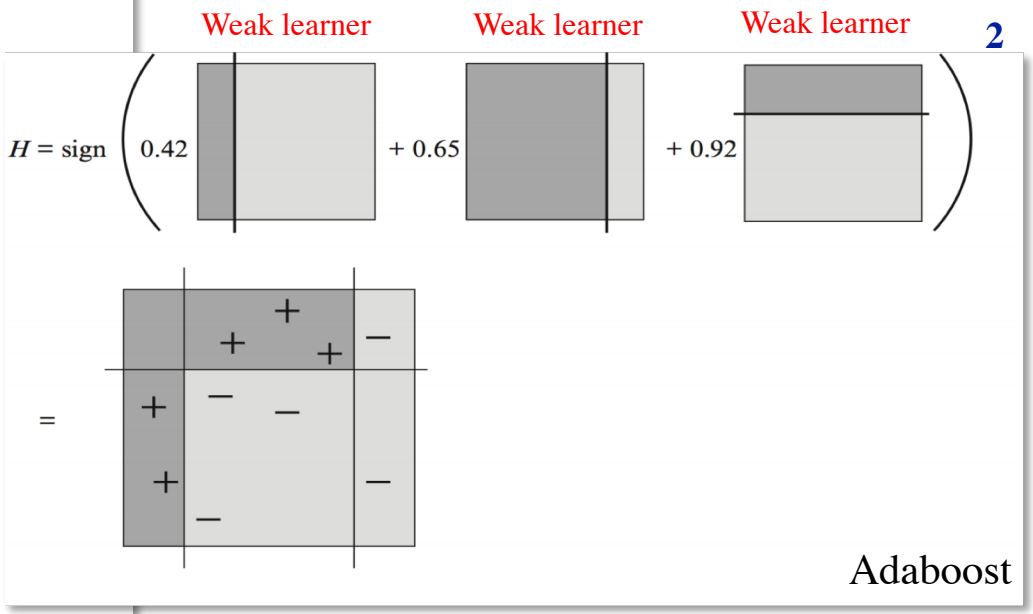
https://youtu.be/0MOH4Dcu5qg

Gradient Boosting – Main Idea

UIC BUSINESS



1



1. Zhang, Zhongxing, et al. "Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning." *Scientific reports*8.1 (2018): 10628.
2. Schapire, Robert E., and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

Consider fitting the following model,

$$f_K(X) = \beta_0 + \beta^\top \Phi(X) = \beta_0 + \sum_{k=1}^{K} \beta_k \phi_k(X)$$

Gradient boosting: start with intercept and expand the model greedily by picking the best function that reduces error most

1. Compute the best intercept,

$$f_0 = \min_{\beta_0} \sum_{i=1}^{N} L(\hat{y}_i, \beta_0)$$

2. Use previous model and compute the next best model,

$$f_k = \min_{\beta_k \varphi_k(\cdot)} \sum_{i=1}^{N} L(\hat{y}_i, f_{k-1} + \beta_k \varphi_k(\hat{x}_i))$$

Numerical Study

UIC BUSINESS

Cohort	Sample size		Outcome (%)			AUROC		
	Study	Repro.	Study	Repro.	Model	Study	GB	LR
Caballero Barajas and Akella (2015), W=24	11,648	11,648	-	13.01	NonLin	0.8657	0.906	0.88616
Caballero Barajas and Akella (2015), W=48	11,648	11,648	-	13.01	NonLin	0.7985	0.9227	0.9034
Caballero Barajas and Akella (2015), W=72	11,648	11,648	-	13.01	NonLin	0.7385	0.9314	0.9144
Calvert et al. (2016b)	3,054	1,985	12.84	13.8	NonLin	0.934	0.9565	0.9025
Calvert et al. (2016a)	9,683	18,396	10.68	14.71	NonLin	0.88	0.9333	0.9110
Celi et al. (2012), AKI	1,400	4,741	30.7	23.92	NonLin	0.875	0.8812	0.8706
Celi et al. (2012), SAH	223	350	25.6	24.86	NonLin	0.958	0.8929	0.8289
Che et al. (2016) (b)	4,000	4,000	13.85	14.35	NonLin	0.8424	0.8461	0.8273
Ding et al. (2016)	4,000	4,000	13.85	14.35	NonLin	0.8177	0.8461	0.8273
Ghassemi et al. (2014), W=12	19,308	28,172	10.84	12.2	Lin	0.84	0.8846	0.8609
Ghassemi et al. (2014), W=24	19,308	23,442	10.80	12.92	Lin	0.841	0.8841	0.8651
Ghassemi et al. (2015)	10,202	21,969	-	13.51	NonLin	0.812	0.8781	0.8591
Grnarova et al. (2016)	31,244	29,572	13.82	12.49	NonLin	0.963	0.9819	0.9765
Harutyunyan et al. (2017)	42,276	45,493	-	10.54	NonLin	0.8625	0.9406	0.9286
Hoogendoorn et al. (2016)	13,923	17,545	-	14.97	NonLin	0.841	0.8797	0.8618
Johnson et al. (2012)	4,000	4,000	-	14.35	NonLin	0.8602	0.8461	0.8273
Johnson et al. (2014)	4,000	4,000	-	14.35	Lin	0.8457	0.8461	0.8273
Joshi and Szolovits (2012)	10,066	10,696	12.0	4.14	Lin	0.89	0.8872	0.8716
Lee and Maslove (2017)	17,490	20,961	17.73	17.86	Lin	0.775	0.8655	0.8488
Lehman et al. (2012)	14,739	21,738	14.6	12.32	Lin	0.82	0.888	0.8694
Pirracchio et al. (2015)	24,508	28,795	12.2	12.72	NonLin	0.88	0.9070	0.8897
Ripoll et al. (2014)	2,002	2,251	21.10	39.63	NonLin	0.8223	0.7900	0.7647

Break

Shopper Marketing Optimization

Shopper Marketing I/III

UIC BUSINESS

- Shopper marketing involves understanding how shoppers behave and react to various marketing tactics and leveraging this to benefit marketers and retailers.
- It is one of the fastest growing forms of marketing for consumer packed goods.
- Some examples:
  - Endcap display
  - Coupons
  - Product display
  - Social media campaign
  - …

Shopper Marketing II/III

UIC BUSINESS

SM considers entire path to purchase to provide incentive to a costumer in the entire path.

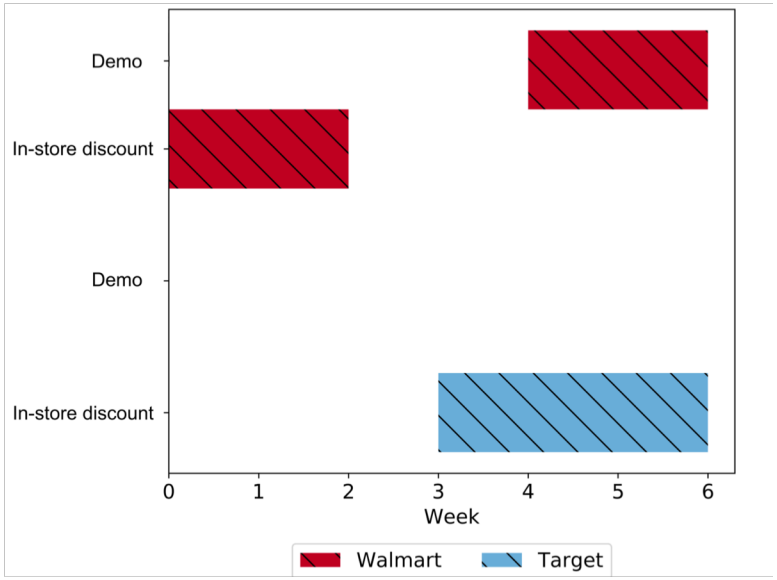
1

1. http://cadentcg.com/what-we-do/our-capabilities/path-to-purchase-research/

Shopper Marketing III/III

UIC BUSINESS

- The core goal of brands employing SM is to choose SM tactics to maximize sales of a product at one or more retailers over a finite planning horizon.

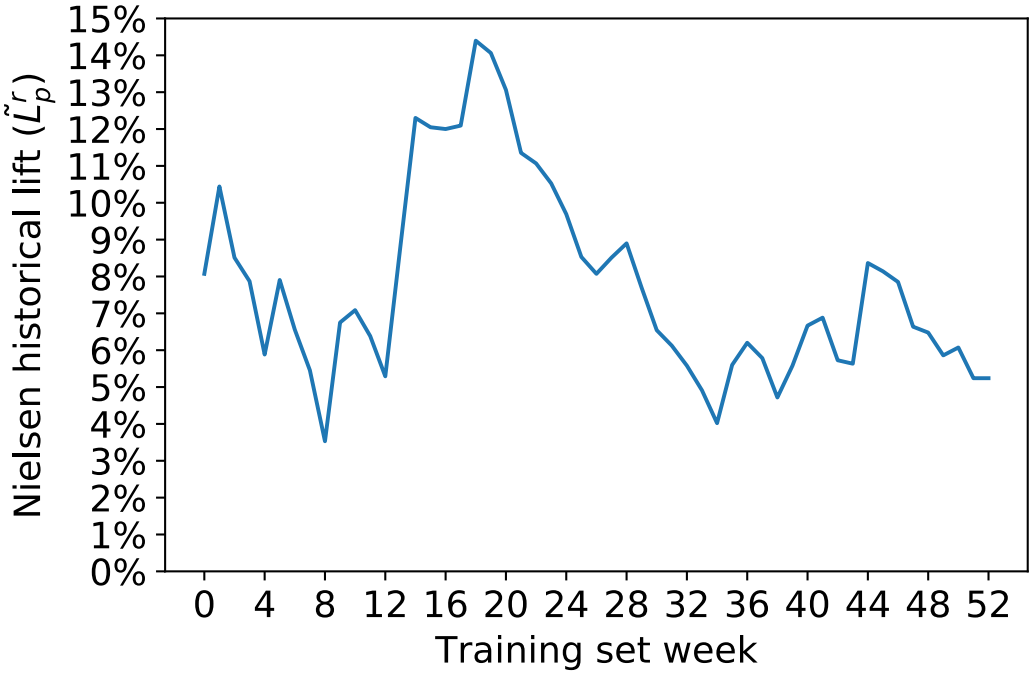


- SM goal can be obtained in two steps:
  1. Lift attribution
  2. Tactic planning

Lift

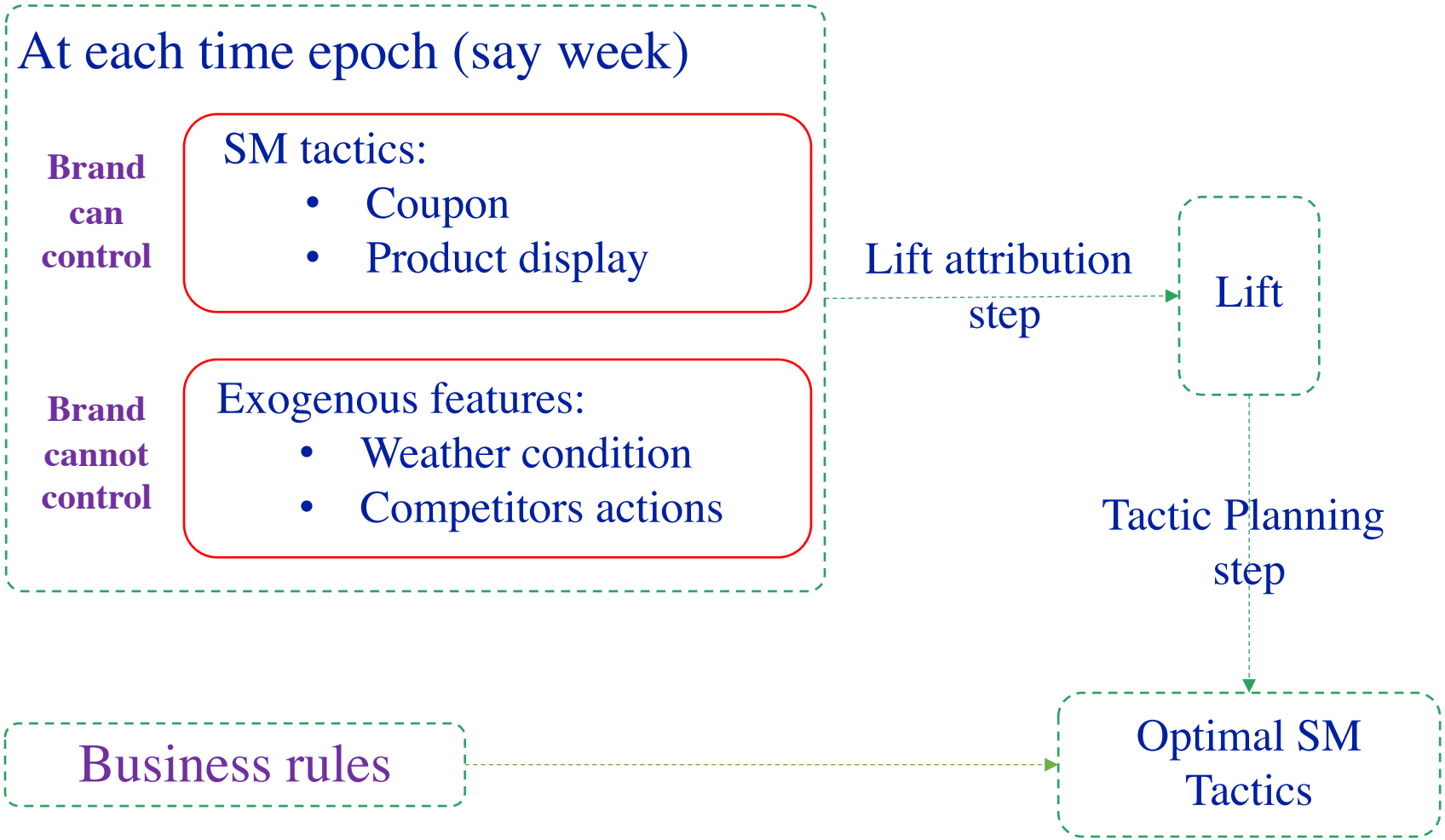
UIC BUSINESS

- Lift: percentage of increment in the volume of sales due to SM tactics.
- A lift predictive model must be able to attribute a component of lift to individual SM tactics.



Lift Attribution & Tactic Planning

UIC BUSINESS



Data Sources

UIC BUSINESS

Historical weekly lift and SM tactics for each retailer

SM tactics

Historical SM tactics in each week

Exogenous features

There might be some historical data

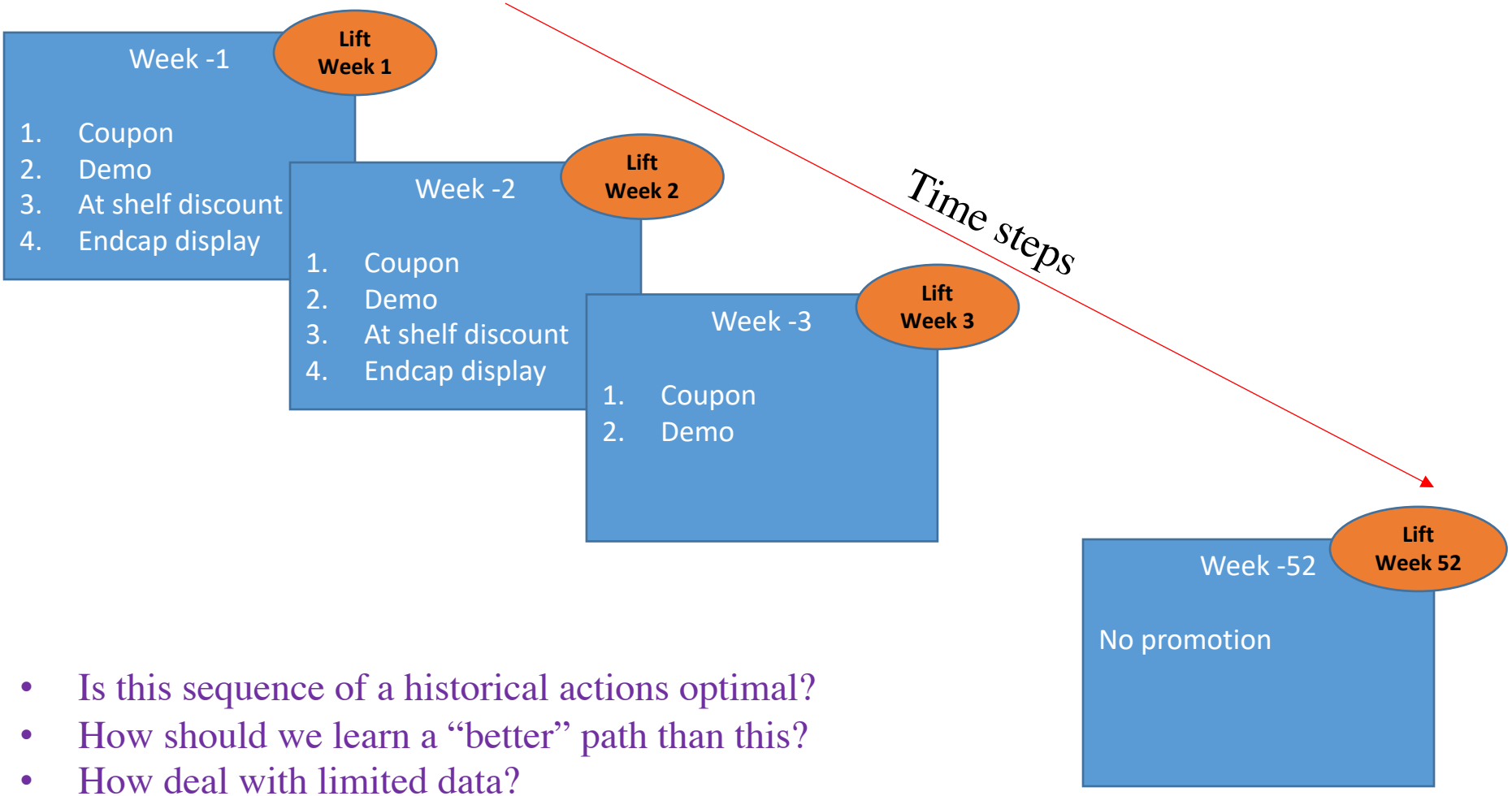
Lift

Sales analytic vendors (Nielsen, IRI, …) estimate POS and weekly lift

- One of the main challenges is that historical lift due to each SM tactic is not available. (think about asking how much increment of sales can be attained by distributing coupon for a week!)
- Data is limited and one should **generalize** a historical sub-optimal path.

Learning from a Sub-optimal Sample Path

UIC BUSINESS



- Is this sequence of a historical actions optimal?
- How should we learn a "better" path than this?
- How deal with limited data?
- How separate SM and non-SM lifts?

An industrial model: Marketing Mix Model (MMM)

1. Regress historical lift versus historical SM-tactics and historical exogenous predictor variables.
2. Cross-validate possibly a huge array of parameters on a validation set.
3. Use a lift model to compute "good" SM-tactics

- For any store dataset, iterate over the above steps until obtaining an "acceptable" model.
- Adding a large array of constraints into these models is either not possible or makes the fitting process expensive.

An industrial model: Marketing Mix Model (MMM)

UIC BUSINESS

- One popular MMM is Bayesian linear mix models to estimate lift i.e.

$$\text{Lift} = \mathcal{N}\Big(\beta^\top(\text{SM-tactics}) + \gamma^\top(\text{Exogenous features}),\ \Sigma\Big)$$

- These models are flexible in handling a large array of exogenous information.

- Some limitations of these models:

  1. They cannot capture temporal effect of SM-planning process
  2. They do not explicitly couple historical SM-tactics with historical lift
  3. They are limited in incorporating business rules (i.e. spending constraints, number of active tactics, …) into a model
  4. These models are strongly data specific
  5. Huge hand engineering is needed

- These issue translate into significant hand engineering of MMMs is needed for obtaining an acceptable lift model, and they could be poor models.

An industrial model: Marketing Mix Model (MMM)

Let's see some recent methods used in industry …

NIELSEN MARKETING MIX MODELING

https://youtu.be/8d16VIq_oJo

Thank You!

Please fill out the feedback form!