

Maximizing ROI

Using State-of-the-Art Data Science Techniques

April 28, 2015

Outline

- ▶ What is ROI?
- ▶ The Walmart problem
- ▶ Data science techniques
- ▶ Conclusions

What is ROI?

- ▶ **Definition:** A profitability measure that evaluates the performance of a business by dividing net profit by cost of investment.

$$\frac{\text{Net Profit}}{\text{Investment}} = \frac{\text{Return} - \text{Investment}}{\text{Investment}}$$

$$= \frac{400,000 - 300,000}{300,000} = 33\frac{1}{3}\%$$

- ▶ **Procedure to improve ROI:**
 - Define “return”
 - Calculate your current return
 - Increase net profit
 - Re-evaluate your expectations

Understanding the problem

- ▶ Walmart store sales data from 2011-2012
- ▶ Five different markdown promotions throughout the year
- ▶ GOAL: Use weekly sales data to optimize promotions and maximize revenue
 - ▶ Which markdowns are worthwhile?
 - ▶ What values of these markdowns are optimal?
 - ▶ When should these markdowns be promoted?

Walmart data

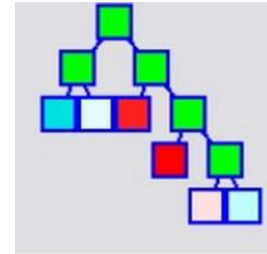
Store	Date	Department	Weekly Sales	Is Holiday	Type	Size	Temperature	Fuel Price	Markdown	CPI	Unemp.
42	9/7/2012	82	7816.9	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	83	3966.1	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	85	14.77	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	87	1774.55	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	90	62244.42	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	91	47899.97	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	92	98189.06	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	93	40510.59	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17
42	9/7/2012	94	41239.14	TRUE	C	39690	83.07	4.124	341.05	130.9325	7.17

Data science techniques

- ▶ Traditional linear regression may be the first step in the modeling process
- ▶ However, regression cannot handle:
 - ▶ Missing values
 - ▶ Nonlinearities
 - ▶ Interactions
 - ▶ Variable selection
- ▶ Instead, we will look at two modern data science techniques to address these problems:
 - ▶ CART decision trees
 - ▶ TreeNet gradient boosting

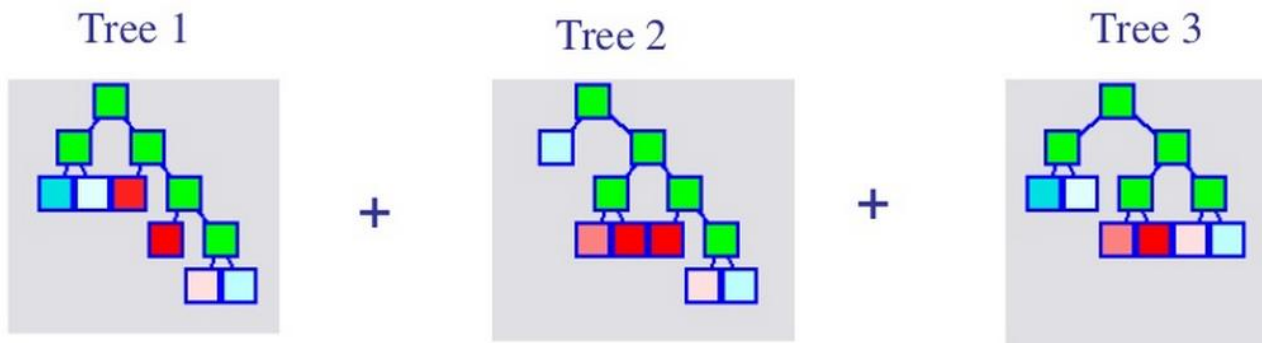
CART

- ▶ Classification and Regression Trees
 - ▶ Separates relevant from irrelevant predictors
 - ▶ Yields simply, easy to understand results
 - ▶ Doesn't require variable transformations
 - ▶ Impervious to outliers and missing values
- ▶ Fastest, most versatile predictive modeling algorithm available to analysts
- ▶ Provides the foundation to modern data mining techniques such as bagging and boosting

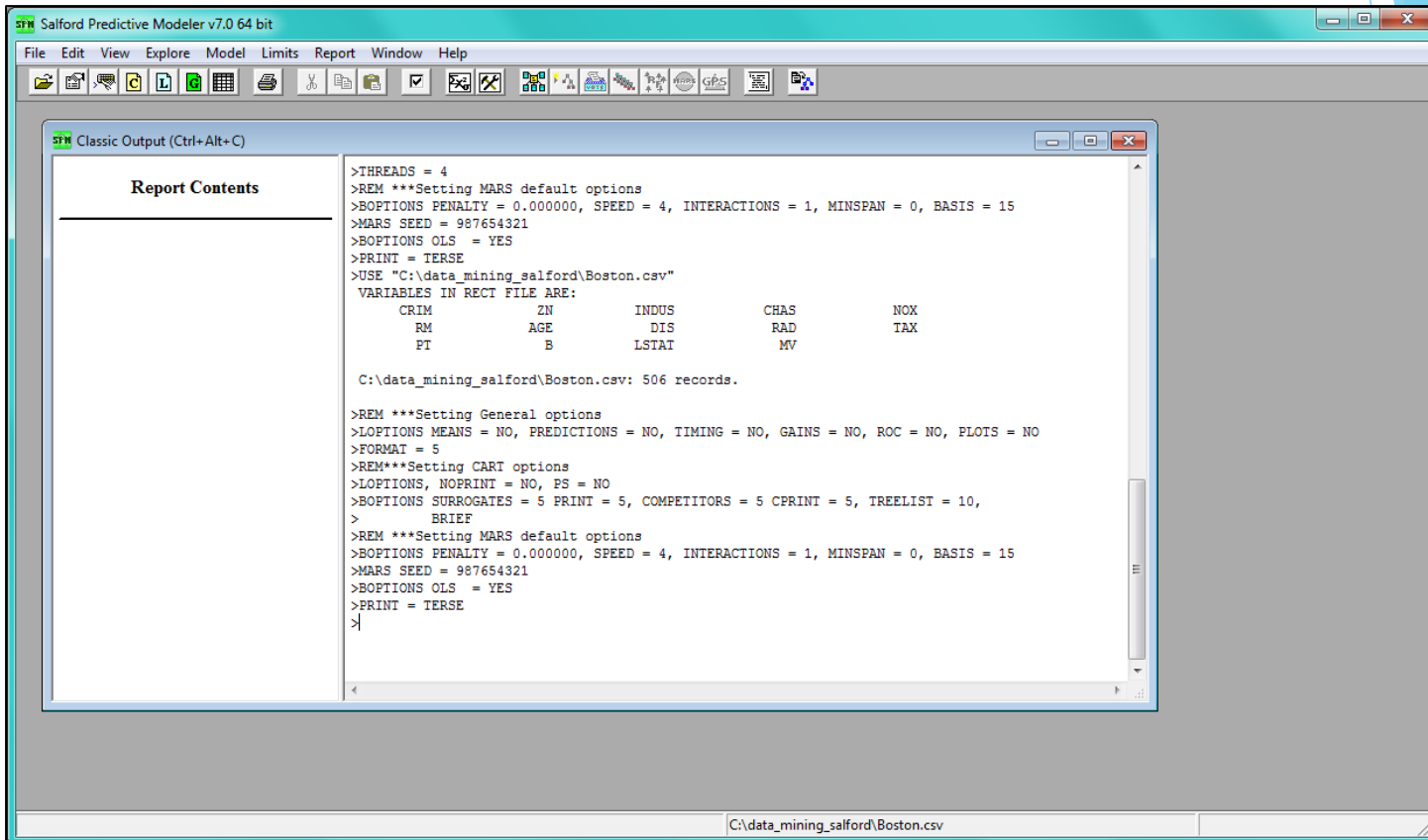


TreeNet

- ▶ Stochastic Gradient Boosting
- ▶ Small decision trees built in an error-correcting sequence
 1. Begin with small tree as initial model
 2. Compute residuals from this model for all records
 3. Grow a second small tree to predict these residuals
 4. And so on...



Building a model



The screenshot shows the Salford Predictive Modeler v7.0 64 bit application window. The main window has a menu bar (File, Edit, View, Explore, Model, Limits, Report, Window, Help) and a toolbar. A 'Classic Output (Ctrl+Alt+C)' window is open, displaying the following text:

```
>THREADS = 4
>REM ***Setting MARS default options
>OPTIONS PENALTY = 0.000000, SPEED = 4, INTERACTIONS = 1, MINSPAN = 0, BASIS = 15
>MARS SEED = 987654321
>OPTIONS OLS = YES
>PRINT = TERSE
>USE "C:\data_mining_salford\Boston.csv"
VARIABLES IN RECT FILE ARE:
      CRIM      ZN      INDUS      CHAS      NOX
      RM      AGE      DIS      RAD      TAX
      PT      B      LSTAT      MV

C:\data_mining_salford\Boston.csv: 506 records.

>REM ***Setting General options
>LOPTIONS MEANS = NO, PREDICTIONS = NO, TIMING = NO, GAINS = NO, ROC = NO, PLOTS = NO
>FORMAT = 5
>REM***Setting CART options
>LOPTIONS, NOPRINT = NO, PS = NO
>OPTIONS SURROGATES = 5 PRINT = 5, COMPETITORS = 5 CPRINT = 5, TREELIST = 10,
>    BRIEF
>REM ***Setting MARS default options
>OPTIONS PENALTY = 0.000000, SPEED = 4, INTERACTIONS = 1, MINSPAN = 0, BASIS = 15
>MARS SEED = 987654321
>OPTIONS OLS = YES
>PRINT = TERSE
>|
```

The status bar at the bottom of the application window shows the file path: C:\data_mining_salford\Boston.csv.

Influential Markdowns

STORE	DEPT	DATE	WEEKLY_SALES	DIFFERENCE_RATE	DIFFERENCE	RESPONSE.a	MARKDOWN3.a	RESPONSE.b	MARKDOWN3.b
45	1	12/23/2011	53942.21	1.989	36952.258	55534.05	1962.19	18581.79	0
25	1	12/23/2011	53439.71	1.934	38389.2892	58237.9	1689.22	19848.61	0
24	1	12/23/2011	49692.85	1.613	31506.922	51039.3	1633.67	19532.37	0
35	1	12/23/2011	37706.15	1.572	26009.4305	42557.3	1429.48	16547.87	0
16	1	12/2/2011	11075.5	1.407	15717.2215	26885.89	944.52	11168.67	0
18	1	12/23/2011	63868.79	1.329	37871.5417	66363.26	2083.28	28491.72	0
19	1	12/23/2011	66366.16	1.289	36505.0922	64826.1	1902.55	28321.01	0
14	1	12/23/2011	101846.1	1.198	38311.0907	70284.09	2127.77	31973	0
23	1	12/23/2011	114300.6	1.044	37025.0772	72498.32	1683.17	35473.25	0
27	1	12/23/2011	78431.1	1.032	29698.0321	58487.97	1438.33	28789.94	0
39	1	12/23/2011	51349.28	1.012	25832.6383	51351.53	1339.1	25518.89	0
20	1	12/23/2011	93745.04	0.99	38090.2178	76558.66	2145.46	38468.44	0
10	1	12/23/2011	134217.7	0.978	45116.2228	91260.19	2432.01	46143.97	0
15	1	12/23/2011	43921.97	0.973	22580.9871	45781.64	1120.42	23200.65	0
22	1	12/23/2011	60924.13	0.956	25365.781	51889.72	936.62	26523.94	0
8	1	12/23/2011	36655.67	0.896	18440.7603	39011.4	891.94	20570.64	0
20	1	12/16/2011	69342.61	0.883	32295.8378	68866.35	1776.71	36570.51	0
4	1	12/23/2011	91965.85	0.79	41576.3282	94204.65	2938.24	52628.32	0
10	1	12/16/2011	100775.8	0.769	35863.3373	82498.97	1203.16	46635.63	0
26	1	12/23/2011	43854.87	0.757	19493.1152	45244.53	1018.94	25751.42	0

$$DIFFERENCE_RATE = \frac{DIFFERENCE}{RESPONSE.Y}$$

Non-influential Markdowns

STORE	DEPT	DATE	WEEKLY_SALES	DIFFERENCE_RATE	DIFFERENCE	RESPONSE.a	MARKDOWN3.a	RESPONSE.b	MARKDOWN3.b
7	1	6/8/2012	6971.47	0	0	7789.085	0.12	7789.085	0
7	1	6/15/2012	7748.14	0	0	8007.492	0.65	8007.492	0
7	1	6/22/2012	7967.26	0	0	8594.52	0.09	8594.52	0
7	1	7/13/2012	9253.35	0	0	9099.11	0.35	9099.11	0
8	1	2/24/2012	13457.69	0	0	14420.63	1	14420.63	0
8	1	3/2/2012	12947.8	0	0	12870.26	0.6	12870.26	0
8	1	9/28/2012	14572.76	0	0	10041.95	0.55	10041.95	0
9	1	2/24/2012	10386.62	0	0	8855.457	1	8855.457	0
9	1	3/23/2012	11993.86	0	0	12085.34	0.68	12085.34	0
9	1	6/8/2012	9289.55	0	0	9155.973	0.3	9155.973	0
9	1	8/3/2012	8765.82	0	0	9280.226	0.22	9280.226	0
9	1	9/28/2012	13772.7	0	0	7266.772	0.55	7266.772	0
19	1	12/30/2011	22068.26	-0.001	-12.3186589	21909.38	210.34	21921.69	0
19	1	7/13/2012	15529.77	-0.001	-9.94868774	14744.39	35.88	14754.34	0
2	1	6/22/2012	24155.67	-0.001	-30.7257629	24756.04	60.99	24786.76	0
2	1	8/3/2012	22851.99	-0.001	-18.8833672	22394.65	43.02	22413.53	0
21	1	1/20/2012	13821.16	-0.001	-13.3913952	14075.11	3.46	14088.5	0
21	1	7/27/2012	10630.35	-0.001	-13.3913952	11127.64	3.43	11141.03	0
5	1	7/20/2012	9513.66	-0.001	-13.3913952	9858.215	3.39	9871.606	0
7	1	8/10/2012	8698.9	-0.001	-13.5278909	10811.5	55.79	10825.03	0
8	1	7/27/2012	9322.04	-0.001	-13.3913952	9288.481	3.55	9301.872	0
1	1	9/7/2012	18322.37	-0.002	-45.9030979	19789.22	50.94	19835.12	0
13	1	8/17/2012	35524.12	-0.002	-65.8642547	33590.01	40.96	33655.88	0

Conclusions

- ▶ Markdowns (promotions) have different effects on sales based on value, time of the year, and other factors.
- ▶ Traditional linear regression cannot capture the nonlinearities and interactions as well as modern data science techniques.
- ▶ Store managers can decide markdown value and when to run the promotion based on this model.
- ▶ Insufficient markdown data may result in an unreliable model for future predictions. You may need more than one year of observations.
- ▶ We cannot describe markdown influence in dollar amounts because values are anonymized, but managers with this knowledge can easily do so.

Other applications

- ▶ **Epidemiology:** intervening in the spread of disease to decrease economic impact
- ▶ **Marketing:** identifying which customers to target in direct-mail marketing to increase response rate
- ▶ **Insurance:** intelligently decreasing (or increasing) premiums to boost total sales
- ▶ **Manufacturing:** optimize process cycle time to decrease costs (labor, machine usage) and increase annual production

Any Questions?

- ▶ Download SPM
 - ▶ 30-day free trial
- ▶ Step-by-step tutorial
- ▶ Data Science Events
 - ▶ Full/Mini-trainings
 - ▶ Upcoming Webinars
- ▶ support@salford-systems.com