Evaluating Stochastic Cost-Benefit Classification Measures For A Retailer's Assortment Mix Decision



Matthew A. Lanham, CAP (lanham@vt.edu)
Doctoral Candidate
MatthewALanham.com

Ralph D. Badinelli, Ph.D. (ralphb@vt.edu)
Lenz Professor of Business
Department of Business Information Technology

VirginiaTech
Pamplin College of Business

Outline

- Assortment Planning
- Demand Modeling
- Binary Classification Modeling
- Class Imbalance
- Model Assessment
- Stochastic Cost-Benefit Classification
- Research Design
- Results
- Future Research

Copyright © 2015, MatthewALanham.com

The Assortment Problem

The Assortment Decision Entails:

- What products to offer where?

Typical Objective:

- Determine which products to stock in a location in order to maximizes sales or profit while satisfying various constraints (e.g. Classical Knapsack Problem)

Importance:

- Considered to be the most important decisions faced by retailers because of the substantial impact the products carried, not carried, or not available have on sales and gross margin (Kök et al., 2015; Sauré & Zeevi, 2013).
- A retailer not providing what the consumers desires will lose market share
- Better assortments can create competitive advantages

Difficulty:

- Thousands of Products & Stores NP-hard Problem
- Understanding and modeling consumer purchase behavior adequately is the holy grail.

The Literature

- Mostly focused on one assortment for a retailer, even though retailers will regularly have different assortments for different stores due to the differences in customer preference.
- It might be assumed that their solutions would then be implemented to each store individually (Kök, Fisher et al. 2015).

Copyright © 2015, MatthewALanham.com

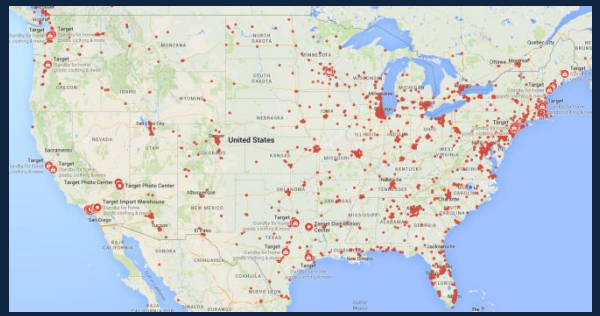
Customized Assortment Problem = Big Data Problem

Store-SKU complexity

- The set of potential SKUs that a retailer may choose from can be 1,000,000+
- A traditional merchandise store (e.g. Big K) would carry approximately 16,000 unique stock-keeping units (SKUs) with an additional 6,000 seasonal products, depending on the time of year (Cox 2011).



- A chain retailer can have several thousands of stores. A few examples, Walmart has 5,187 stores in the U.S. alone (Walmart 2015), Target 2,155 (Target 2015), and J.C. Penny 1,060 (Wikipedia 2015).



- How can we possibly model demand so that we know which SKUs to put in which stores?

Copyright © 2015, MatthewALanham.com

Demand Modeling

Focused on Long-term Planning Horizon Products

- Service parts
- Office products



Forecasting Techniques

- ETS models, ARIMA, etc.
- Goal: Estimate number of demand units for next t periods
- Forecasting time-series with "lumpy" demand if very difficult

Binary Classification Approaches

- Logistic Regression, Decision Trees, Neural Networks, etc.
- Goal: Predict purchase propensity of a product if stocked in a respective store

Multi-Class Classification Approaches

- Multinomial Logit Model (MNL), Nested Logit (NL), Locational-choice (LC), Exogenous Demand (ED)
- Goal: Predictive purchase propensity and estimate substitution behavior within a set of similar products (i.e. substitutes) for a respective store
- Academic literature has primarily focused on these utility-based models to estimate substitution effects

Copyright © 2015, MatthewALanham.com

Binary Classification Modeling

Binary Classification

Objective: Determine the probability that SKU *i* will sell in store *j*

- This is not being performed in the academic literature for various reasons (type of retail business, product turn rates, etc.) but it can be effective for long-term planning horizon products
- We have found that if one is to use this approach selecting SKU propensity to sell measures based on traditional predictive model performance measures (e.g. ROC/AUC) is insufficient

Set up

- Binary classification algorithms fit a model having a binary response (e.g. No/Yes, 0/1, etc.).
- When the model is a fit, a continuous prediction for each record is generated that lies between 0 and 1.
- These predictions are then compared to a specified decision cutoff criterion, which allows each prediction to be classified as a member of only one of two possible disjoint classes.
- Our binary classification models are employed on previously stocked stock-keeping-unit (SKU) that have sold over a particular time interval in any store.
- Each SKU that has sold over a specified number of units as deemed by the retailer are considered “sellers” (or Class 1), while those that have not sold more than a specified number of units over a certain time interval are labeled as “non-sellers” (Class 0).

Copyright © 2015, MatthewALanham.com

Class Imbalance

- Class imbalance occurs when one class has more observations than another class. For example, training a model with 100 records where 70 had a response value of “Class 1”, while the remaining records had a response value of “Class 0”, which would yield a minority class imbalance of 30%.
- Studies have shown that not taking into account the class imbalance can have negative consequences on model estimation [24]. Essentially, the imbalance makes it difficult for the binary classification to learn, which usually leads to correctly classifying majority class records and not the minority class records [25].
- Today, there is not one technique that works optimally well for all data sets and researchers continue to work on creating a unified rebalancing framework to tackle this problem [26]. He and Garcia [27] provide a throughout review of this area. In our study, we investigate non-balancing, down-Sampling, up-Sampling, SMOTE, and ROSE.
- In this study we investigate the combination of different binary classification algorithms and class balance approaches:
  - Non-balancing
  - Down-sampling
  - Up-sampling
  - Synthetic Minority Over-sampling Technique (SMOTE)
  - Random OverSampling Examples (ROSE)

Copyright © 2015, MatthewALanham.com

Model Assessment

- When predictive models are constructed they must be empirically validated. Typically, the process involves fitting/training a model on one data set, and then once the model is constructed to feed in another data set that was not used to train the model commonly referred to as a testing set.

Training dataset

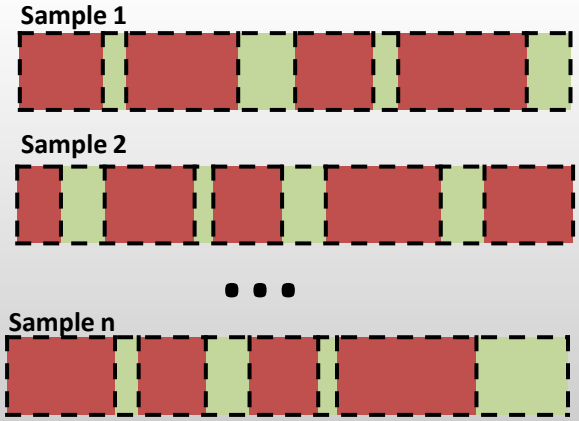
Testing dataset

This testing dataset is never used (or looked at) for cross-validation below. It is used at the very end when are ready to assess the "true" performance of your model once.

A few common cross-validation techniques to be performed explicitly on your training data set ONLY.

Random subsampling

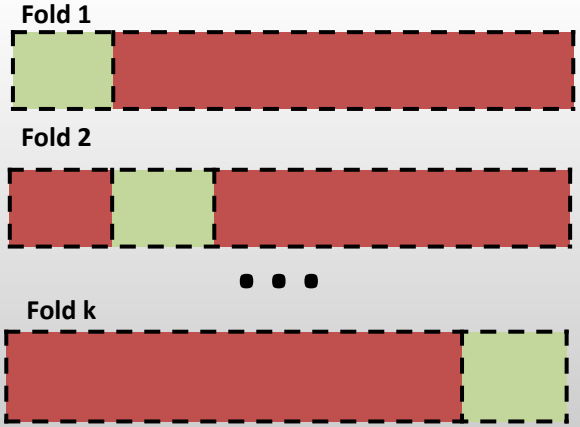
1) Build and test models for each sample



2) Average over all samples you generated to get an estimate of the expected test error rate

k-fold

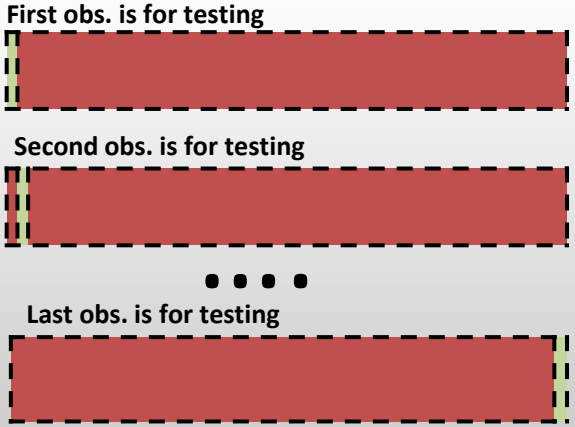
1) Partition data into equal-sized disjoint subsets and build and test each fold



2) Average over folds you generated to get an estimate of the expected test error rate

Leave-one-out (LOO)

1) Partition data into subsets so as to leave exactly 1 element out for testing



2) Average over all to get an estimate of the expected test error rate

- This cross-validation procedure provides a proxy of truth about the validity of the model as well as gauges the expected performance of future predictions when new data are used.

Copyright © 2015, MatthewALanham.com

Confusion Matrix

- The confusion matrix is a cross-tabulation table that provides a gauge of how well a model's predictions were classified compared to the response. Since the response only has two possible classes in our case (i.e. sold vs. not sold), our predictions, in the form of a probability between $[0,1]$, are assigned to one of these potential classes.

		Observed (Y)		
		Sell	Not Sell	
		1	0	
Predicted $(\hat{Y})$	1	*TP*	*FP*	*TP+FP*
	0	*FN*	*TN*	*FN+TN*
		TP+FN	*FP+TN*	Total

- The assignment is based on the modeler's specified cutoff threshold, which is most often 0.5. This means that if a prediction realizes a value of 0.67, it will be assigned to the "seller" class or "Class 1" because the prediction is larger than the threshold parameter. Likewise, a value of 0.49 would be assigned to "Class 0" or "non-seller" and so on for all probability predictions.
- A large number of true-positives (TP) and true-negatives (TN) on the diagonal of the matrix and a small number of false-positives (FP) and false-negatives (FN) provides an indication that the model used performs well at identifying SKUs that will sell or not sell in our context.

Copyright © 2015, MatthewALanham.com

Traditional Assessment Statistics from Confusion Matrix

Statistics

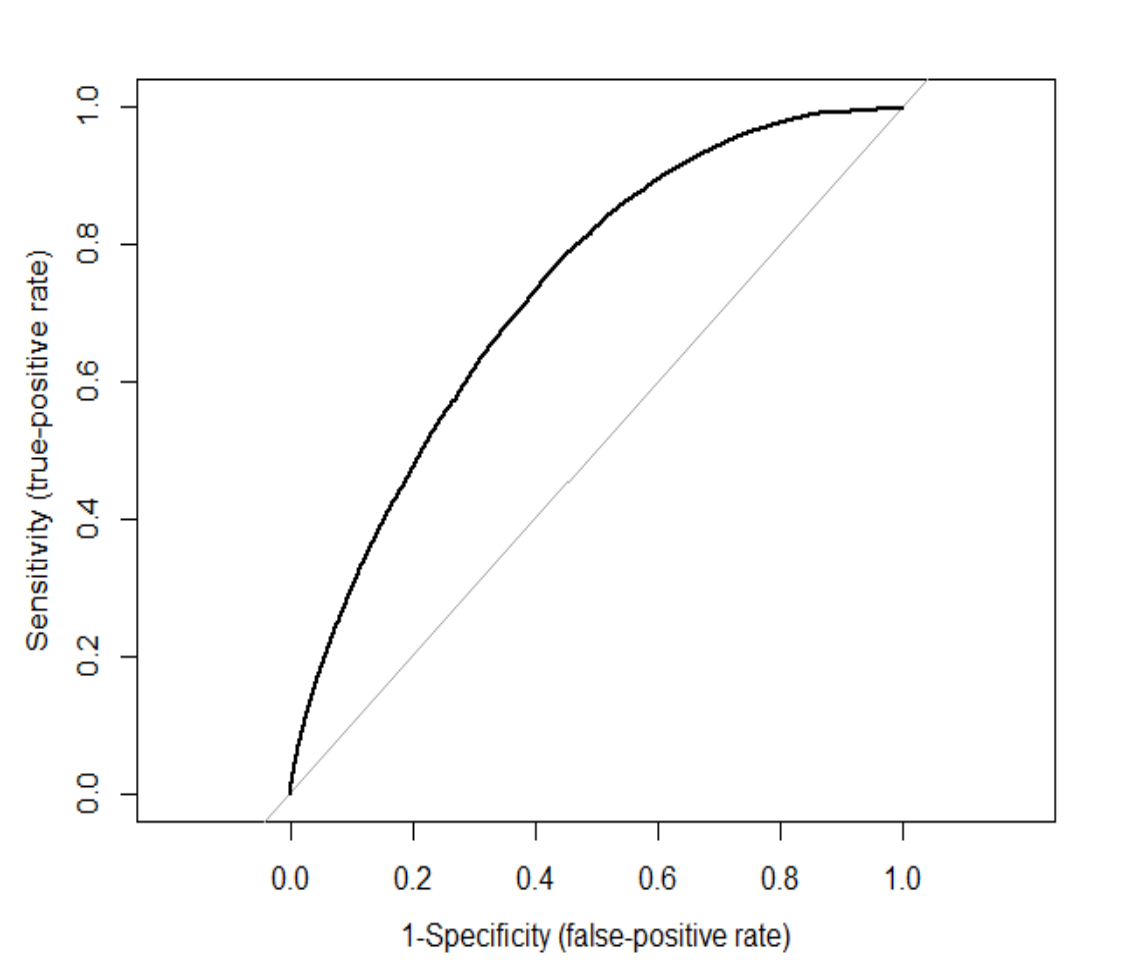
Statistic	Alternative Name	Description	Formula
Overall Accuracy		Probability that a SKU is classified as a seller and non-seller correctly	$\frac{TP+TN}{Total}$
Sensitivity	True-Positive Rate; Recall	Probability that a retailer will predict a sell when there is actually sell	$\frac{TP}{TP+FN}$
Specificity	True-Negative Rate	Probability that a retailer will not predict a sell when there is not a sell	$\frac{TN}{FP+TN}$
Type I error rate	False-Positive Rate; 1- Specificity	Probability that a retailer will predict a sell when there is not a sell	$\frac{FP}{FP+TN}$
Type II error rate	False-Negative Rate	Probability that a retailer will not predict a sell when there is actually sell	$\frac{FN}{TP+FN}$
Positive Predictive Value (PPV)	Precision	Probability that the SKU did sell when it actually sold (an unconditional analog to sensitivity, which takes into account the event's prevalence)	$\frac{TP}{TP+FP}$
Negative Predictive Value (NPV)		Probability that the SKU did not sell when it actually did not sell (an unconditional analog to specificity, which takes into account the event's prevalence)	$\frac{TN}{FN+TN}$
Positive Likelihood Ratio (PLR)		Ratio between the probability that the retailer predicts a sell when there is a sell and the probability that a SKU will sell given it actually did not sell	$\frac{Sensitivity}{1-Specificity}$
Negative Likelihood Ratio (NLR)		Ratio between the probability that a retailer predicts a SKU will not sell given it actually sold and the probability that the retailer predicts that the SKU will not sell given that it did not sell	$\frac{1-Sensitivity}{Specificity}$
Youden's J Index Youden [28]	J Index	The proportion of correctly predicted samples for both the seller and non-seller groups (an alternative to the ROC curve)	Sensitivity + Specificity - 1
Cohen's Kappa	Kappa	Historically been used to assess agreement among two raters but is appropriate to the retailer in this context as well [29]. If we let O = observed accuracy and E = expected accuracy, we can calculate Cohen's Kappa statistic based on the confusion matrices' marginal totals. Let, $O = TP + TN$ and $E = \frac{(TP+FP)*(TP+FN)}{(Total)} + \frac{(FN+TN)*(FP+TP)}{(Total)}$	$\frac{O-E}{Total-E}$

Copyright © 2015, MatthewALanham.com

The ROC Curve

Receiver Operating Characteristic (ROC) curve

- The **ROC curve** is the most frequently used device to compare and assess different models.
- ROC is constructed from confusion matrices with various cutoff thresholds (e.g. 0.001 to 0.999)



Copyright © 2015, MatthewALanham.com

The ROC Curve

Thinking more about the problem…

In the context of our study, the sensitivity statistic estimates the probability that the retailer will predict a seller when the SKU was indeed actually a seller (i.e. sold), so it is a true-positive rate.

A retailer would want sensitivity to be as high as possible, but the higher the sensitivity becomes leads to a negative tradeoff with the false-positive rate, meaning the retailer is more likely to predict and classify a SKU as a seller when in fact it is a non-seller. This is important to the retailer because stocking products that will not sell lead to additional inventory costs and take up shelf space for products that could sell.

Moreover, the longer seasonal effects of having to markdown unwanted products hurts the retailer's margin, which based on the quantity of the markdowns, can lead to dramatic consequences to the retailer's bottom line.

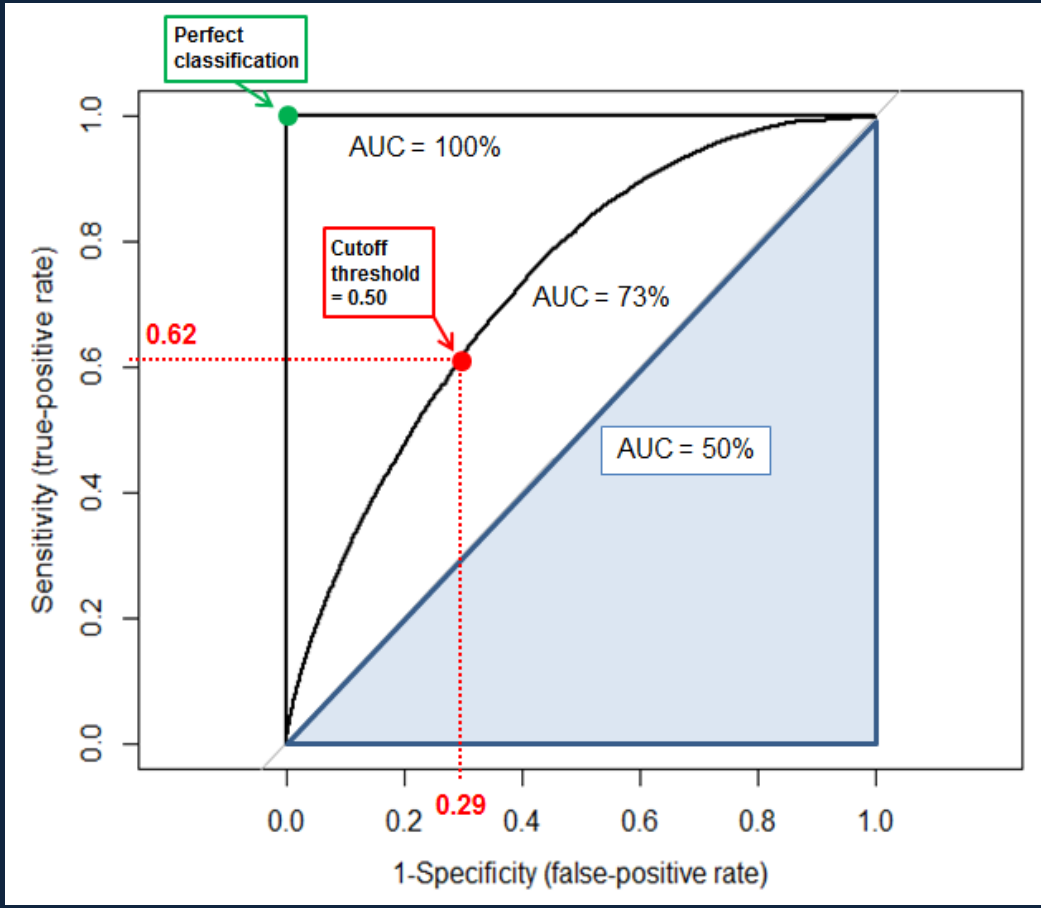


Copyright © 2015, MatthewALanham.com

Area Under The ROC Curve

Area Under the Curve (AUC)

- To assess the overall performance of the model based on the ROC curve, the physical area under the curve (AUC) can be calculated. A curve having no classification ability will lie perfectly on the 45 degree line, thus the AUC is 0.50. Ideally, the retailer would prefer an ROC curve that goes from point (0,0) to (0,1) to (1,1) which would indicate perfect classification for any cutoff threshold and lead to an AUC of 1 as shown in Figure 3.



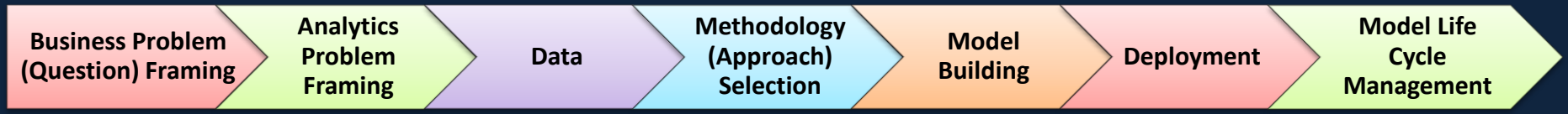
Copyright © 2015, MatthewALanham.com

Stochastic Cost-Benefit Analysis

Motivation

Provost and Fawcett [33], *"Even if a model passes strict evaluation tests "in the lab," there may be external considerations that make it impractical. For example, a common flaw with detection solutions (such as fraud detection, span detection, and intrusion monitoring) is that they produce too many false alarms. A model may be extremely accurate (>99%) by laboratory standards, but* ***evaluation in the actual business context*** *may reveal that it still produces too many false alarms to be economically feasible."*

INFORMS Certified Analytics Professional Domains



- When I was trying to apply the CAP framework to my research, I felt like the connection of the business problem to the analytics problem was disconnected.
- At the end of the day, I want statistically valid model(s) that have performance measures that can be understood by my decision-makers (e.g. Category managers) to help them make better decisions.



Copyright © 2015, MatthewALanham.com

Stochastic Cost-Benefit Analysis

Example

Consider ten products from a similar product category. These products will have similar domain characteristics, but will have varying prices, costs, margins, and profits. We can see that historically certain products have sold over a specified time frame (i.e. "sellers") and some have not sold (i.e. "non-sellers"). The model estimates a probability that each product will sell and based on the cutoff threshold of 0.50 each product can be assigned to a class.

SKU i	Sales	Cost	Profit	Yhat	Yhat[0-1 | Cutoff=0.50]	Y	Expected Return
A	$15.00	$ 9.00	$ 6.00	0.950	1	0	$ (9.00)
B	$ 1.00	$ 0.46	$ 0.54	0.465	1	1	$ 0.54
C	$ 3.56	$ 1.50	$ 2.06	0.829	1	1	$ 2.06
D	$17.28	$10.00	$ 7.28	0.460	0	1	$ (7.28)
E	$ 5.89	$ 2.00	$ 3.89	0.959	1	0	$ (2.00)
F	$ 9.69	$ 5.00	$ 4.69	0.392	0	0	$ -
G	$ 2.45	$ 1.20	$ 1.25	0.460	0	0	$ -
H	$30.00	$15.00	$15.00	0.620	1	1	$ 15.00
I	$10.22	$ 5.62	$ 4.60	0.320	0	0	$ -
J	$ 3.25	$ 1.49	$ 1.76	0.999	1	1	$ 1.76

Copyright © 2015, MatthewALanham.com

Stochastic Cost-Benefit Analysis

Lets do some math…

Below are some basic equations we could us to calculate the expected return for correctly classifying each SKU.

Predicted $(\hat{Y})$ \ Observed (Y)	Sell (1)	Not Sell (0)
1	$\sum_{SKU_i} Rev_i - \sum_{SKU_i} Cost_i$	$-\sum_{SKU_i} Cost_i$
0	$-\left(\sum_{SKU_i} Rev_i - \sum_{SKU_i} Cost_i\right)$	0

A retailer will want to try and minimize the false-positive and false-negatives.

- Some false-negatives will have a lost sales cost of not stocking the product in a location when it would have sold.
- False-positives will have their own costs as well, such as inventory, recycling, and lost opportunity costs because the product was purchased and place in a location and it did not sell.

Copyright © 2015, MatthewALanham.com

Stochastic Cost-Benefit Analysis

In our scenario we only consider the cost of the product and do not incorporate those additional inventory and recycling costs.

In the financial form of classification we show here, if the retailer correctly predicted SKU i would sell they would know that the associated profit for that SKU is

$$p_i = rev_i - cost_i$$

However, if SKU i does not sell but was predicted to have sold they would earn $-cost_i$ for making that stocking decision.

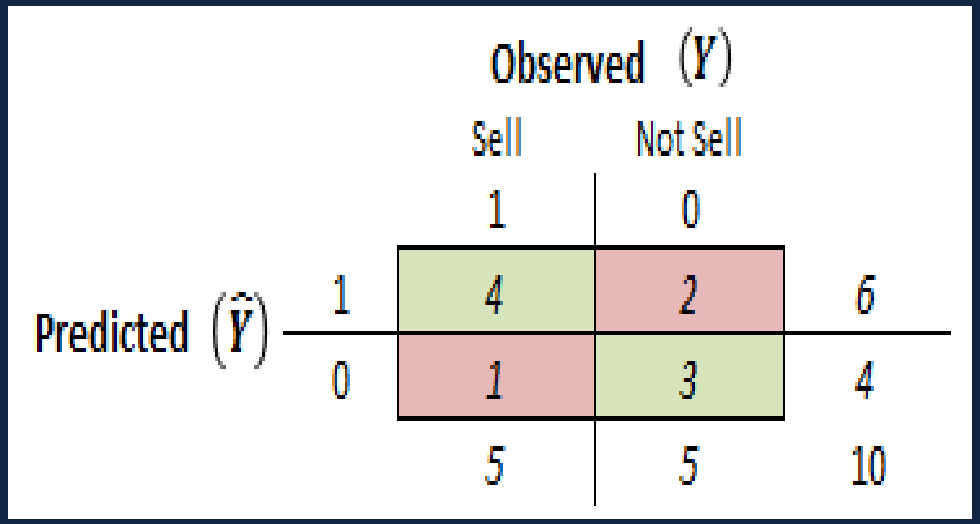
		Observed (Y): Sell (1)	Observed (Y): Not Sell (0)
Predicted $(\hat{Y})$	1	$\sum_{SKU_i} Rev_i - \sum_{SKU_i} Cost_i$	$-\sum_{SKU_i} Cost_i$
	0	$-\left(\sum_{SKU_i} Rev_i - \sum_{SKU_i} Cost_i\right)$	0

Copyright © 2015, MatthewALanham.com

Stochastic Cost-Benefit Analysis

Here is the traditional confusion matrix based on counts of correctly classified and incorrectly classified SKUs from our example. Evaluating the classification accuracy for this model, the overall accuracy is 70% (7/10), sensitivity is 80% (4/5), and specificity is 60% (3/5).

SKU i	Yhat[0-1 | Cutoff=0.50]	Y
A	1	0
B	1	1
C	1	1
D	0	1
E	1	0
F	0	0
G	0	0
H	1	1
I	0	0
J	1	1



Copyright © 2015, MatthewALanham.com

Stochastic Cost-Benefit Analysis

Now, lets look at the financial classification cost-benefit of using this model to stock SKUs classified as sellers or non-sellers. In this example, the model would yield a positive net gain.

SKU i	Expected Return
A	$ (9.00)
B	$ 0.54
C	$ 2.06
D	$ (7.28)
E	$ (2.00)
F	$ -
G	$ -
H	$ 15.00
I	$ -
J	$ 1.76

		Observed (Y): Sell 1	Observed (Y): Not Sell 0	
Predicted $(\hat{Y})$	1	$ 19.36	$ (11.00)	$ 8.36
	0	$ (7.28)	$ -	$ (7.28)
		$ 12.08	$ (11.00)	$ 1.08
			19.36	$ (18.28)

However, the misclassification of SKU A has a major overall impact to the business, possibly leading to an overall loss for this product category because the retailer is not incorporating the additional inventory, recycling, and lost opportunity costs.

Copyright © 2015, MatthewALanham.com

Research Design

To evaluate our proposed stochastic cost-benefit performance measures we tested four different binary classification algorithms (Logistic Regression, Classification tree, C5.0 decision tree, Linear Discriminant Analysis) on one product category from a national retailer.

The dataset was randomly split into a 70% training dataset and a 30% out-of-sample testing dataset. Due to the dataset being imbalanced, the training dataset was rebalanced using four commonly used rebalancing techniques (up-Sampling, down-Sampling, SMOTE, and ROSE) as well as evaluated without rebalancing.

We compare the different modeling approaches to identify what the optimal model chosen would have been based solely on traditional binary classification performance measures (e.g. AUC, Overall accuracy) and then compare those results to the stochastic cost-benefit evaluation procedure to see how well they correlate.

For this one product category, this design allows us to answer these three research questions:

R1: What are the expected business gains a retailer could expect to achieve when using various combinations of rebalancing and binary classification algorithms to identify SKU sellers from non-sellers?

R2: For each rebalance technique/binary classification algorithm combination, can the retailer expect the cost-benefit on the out of sample test set to follow more or less in line with the training data set? Classical statistics do not, as one might have over fit the training dataset.

R3: For each rebalance technique/binary classification algorithm combination, can the retailer expect the cost-benefit on the out of sample test datasets to be similar and could using the financial measures be used to identify which model is optimal compared to traditional statistical measures?

Copyright © 2015, MatthewALanham.com

Results

Training data statistics per rebalance set and model

Rebalance	Model	AUC	Accuracy Pct	TP	FP	FN	TN	Expected Profit	Misclassification Costs
Raw	C5.0	0.7553	83.52 (83.18,83.85)	$ 1,159,602	$ (153,131)	$ (68,296)	$ -	$ 938,174	$ (221,428)
	CART	0.7356	81.48 (81.13,81.83)	$ 1,130,527	$ (176,867)	$ (97,371)	$ -	$ 856,289	$ (274,238)
	LDA	0.7892	77.73 (77.35,78.1)	$ 1,202,326	$ (280,017)	$ (25,571)	$ -	$ 896,738	$ (305,588)
	Logit	0.8103	80.49 (80.13,80.84)	$ 1,137,988	$ (182,159)	$ (89,910)	$ -	$ 865,918	$ (272,069)
	Average	**0.7726**	**80.8 (80.45,81.15)**	**$ 1,157,611**	**$ (198,043)**	**$ (70,287)**	**$ -**	**$ 889,280**	**$ (268,331)**
Down	C5.0	0.8427	77.21 (76.64,77.77)	$ 258,798	$ (47,210)	$ (99,993)	$ -	$ 111,596	$ (147,202)
	CART	0.7660	73.55 (72.95,74.13)	$ 247,208	$ (61,222)	$ (111,583)	$ -	$ 74,404	$ (172,805)
	LDA	0.7949	72.05 (71.45,72.65)	$ 198,095	$ (38,931)	$ (160,696)	$ -	$ (1,531)	$ (199,626)
	Logit	0.8080	73.53 (72.93,74.11)	$ 209,329	$ (42,520)	$ (149,462)	$ -	$ 17,347	$ (191,982)
	Average	**0.8029**	**74.08 (73.49,74.67)**	**$ 228,358**	**$ (47,471)**	**$ (130,433)**	**$ -**	**$ 50,454**	**$ (177,904)**
Up	C5.0	0.9133	85.68 (85.43,85.93)	$ 1,020,521	$ (121,283)	$ (207,377)	$ -	$ 691,860	$ (328,660)
	CART	0.7658	73.63 (73.31,73.94)	$ 844,728	$ (212,999)	$ (383,169)	$ -	$ 248,560	$ (596,168)
	LDA	0.7969	72.33 (72.01,72.65)	$ 683,576	$ (135,308)	$ (544,322)	$ -	$ 3,946	$ (679,630)
	Logit	0.8093	73.46 (73.14,73.78)	$ 715,510	$ (147,392)	$ (512,388)	$ -	$ 55,730	$ (659,780)
	Average	**0.8213**	**76.27 (75.97,76.58)**	**$ 816,084**	**$ (154,245)**	**$ (411,814)**	**$ -**	**$ 250,024**	**$ (566,059)**
SMOTE	C5.0	0.8895	81.44 (81.16,81.72)	$ 1,292,446	$ (275,656)	$ (141,012)	$ -	$ 875,779	$ (416,667)
	CART	0.7409	71.57 (71.25,71.89)	$ 1,073,896	$ (352,472)	$ (359,562)	$ -	$ 361,863	$ (712,033)
	LDA	0.7588	69.61 (69.29,69.94)	$ 1,035,752	$ (284,895)	$ (397,705)	$ -	$ 353,152	$ (682,600)
	Logit	0.7893	73.16 (72.84,73.47)	$ 1,031,881	$ (219,117)	$ (401,577)	$ -	$ 411,187	$ (620,694)
	Average	**0.7946**	**73.95 (73.63,74.26)**	**$ 1,108,494**	**$ (283,035)**	**$ (324,964)**	**$ -**	**$ 500,495**	**$ (607,999)**
ROSE	C5.0	0.9621	91.62 (91.37,91.86)	$ 730,055	$ (40,267)	$ (79,993)	$ -	$ 609,795	$ (120,260)
	CART	0.7437	74.17 (73.78,74.56)	$ 418,419	$ (59,269)	$ (391,630)	$ -	$ (32,480)	$ (450,899)
	LDA	0.7603	70.18 (69.77,70.59)	$ 356,813	$ (69,564)	$ (453,235)	$ -	$ (165,986)	$ (522,799)
	Logit	0.7628	71 (70.6,71.41)	$ 399,103	$ (83,621)	$ (410,945)	$ -	$ (95,463)	$ (494,566)
	Average	**0.8072**	**76.74 (76.38,77.11)**	**$ 476,098**	**$ (63,180)**	**$ (333,951)**	**$ -**	**$ 78,967**	**$ (397,131)**
Overall Average		**0.7997**	**76.37 (75.99,76.75)**	**$ 757,329**	**$ (149,195)**	**$ (254,290)**	**$ -**	**$ 353,844**	**$ (403,485)**

Copyright © 2015, MatthewALanham.com

Results

Training and testing assessment statistics comparison

		Training Data		Testing Data	
Rebalance	Model	AUC	Accuracy Pct	AUC	Accuracy Pct
Raw	C5.0	0.7553	83.52 (83.18,83.85)	0.7561	83.6 (83.09,84.1)
	CART	0.7356	81.48 (81.13,81.83)	0.7383	81.59 (81.05,82.11)
	LDA	0.7892	77.73 (77.35,78.1)	0.7957	77.55 (76.97,78.12)
	Logit	0.8103	80.49 (80.13,80.84)	0.8153	80.68 (80.14,81.22)
	Average	**0.7726**	**80.8 (80.45,81.15)**	**0.7763**	**80.85 (80.31,81.39)**
Down	C5.0	0.8427	77.21 (76.64,77.77)	0.8412	75.27 (74.68,75.86)
	CART	0.7660	73.55 (72.95,74.13)	0.7707	74.55 (73.95,75.14)
	LDA	0.7949	72.05 (71.45,72.65)	0.8014	68.58 (67.94,69.22)
	Logit	0.8080	73.53 (72.93,74.11)	0.8128	70.82 (70.19,71.44)
	Average	**0.8029**	**74.08 (73.49,74.67)**	**0.8065**	**72.31 (71.69,72.91)**
Up	C5.0	0.9133	85.68 (85.43,85.93)	0.9150	85.48 (84.99,85.96)
	CART	0.7658	73.63 (73.31,73.94)	0.7721	74.61 (74.01,75.21)
	LDA	0.7969	72.33 (72.01,72.65)	0.8024	68.61 (67.97,69.25)
	Logit	0.8093	73.46 (73.14,73.78)	0.8136	70.97 (70.34,71.59)
	Average	**0.8213**	**76.27 (75.97,76.58)**	**0.8258**	**74.92 (74.33,75.5)**
SMOTE	C5.0	0.8895	81.44 (81.16,81.72)	0.8546	84 (83.49,84.5)
	CART	0.7409	71.57 (71.25,71.89)	0.7300	77.9 (77.32,78.46)
	LDA	0.7588	69.61 (69.29,69.94)	0.7569	73.16 (72.55,73.76)
	Logit	0.7893	73.16 (72.84,73.47)	0.7866	75.93 (75.34,76.52)
	Average	**0.7946**	**73.95 (73.63,74.26)**	**0.7820**	**77.75 (77.18,78.31)**
ROSE	C5.0	0.9621	91.62 (91.37,91.86)	0.7299	35.06 (34.41,35.71)
	CART	0.7437	74.17 (73.78,74.56)	0.6543	51.15 (50.47,51.84)
	LDA	0.7603	70.18 (69.77,70.59)	0.7995	63.78 (63.12,64.44)
	Logit	0.7628	71 (70.6,71.41)	0.8042	66.77 (66.12,67.41)
	Average	**0.8072**	**76.74 (76.38,77.11)**	**0.7470**	**54.19 (53.53,54.85)**
Overall Average		**0.7997**	**76.37 (75.99,76.75)**	**0.7875**	**72 (71.41,72.59)**

Copyright © 2015, MatthewALanham.com

Testing Data Statistics

Our results show that the financial performance measures follow closely in line with the statistical measures.

Rebalance	Model	AUC	Accuracy Pct	TP	FP	FN	TN	Expected Profit	Misclassification Costs
Raw	C5.0	0.7561	83.6 (83.09,84.1)	$ 486,671	$ (63,688)	$ (29,452)	$ -	$ 393,531	$ (93,140)
	CART	0.7383	81.59 (81.05,82.11)	$ 470,624	$ (73,996)	$ (45,499)	$ -	$ 351,129	$ (119,495)
	LDA	0.7957	77.55 (76.97,78.12)	$ 504,976	$ (120,780)	$ (11,147)	$ -	$ 373,049	$ (131,927)
	Logit	0.8153	80.68 (80.14,81.22)	$ 477,356	$ (77,276)	$ (38,767)	$ -	$ 361,313	$ (116,043)
	Average	**0.7763**	**80.85 (80.31,81.39)**	**$ 484,907**	**$ (83,935)**	**$ (31,216)**	**$ -**	**$ 369,756**	**$ (115,151)**
Down	C5.0	0.8412	75.27 (74.68,75.86)	$ 369,293	$ (18,835)	$ (146,830)	$ -	$ 203,627	$ (165,665)
	CART	0.7707	74.55 (73.95,75.14)	$ 356,504	$ (25,467)	$ (159,619)	$ -	$ 171,418	$ (185,086)
	LDA	0.8014	68.58 (67.94,69.22)	$ 290,155	$ (16,058)	$ (225,968)	$ -	$ 48,129	$ (242,026)
	Logit	0.8128	70.82 (70.19,71.44)	$ 303,645	$ (17,835)	$ (212,478)	$ -	$ 73,331	$ (230,314)
	Average	**0.8065**	**72.31 (71.69,72.91)**	**$ 329,899**	**$ (19,549)**	**$ (186,224)**	**$ -**	**$ 124,126**	**$ (205,773)**
Up	C5.0	0.9150	85.48 (84.99,85.96)	$ 430,926	$ (14,352)	$ (85,197)	$ -	$ 331,378	$ (99,549)
	CART	0.7721	74.61 (74.01,75.21)	$ 356,921	$ (25,506)	$ (159,203)	$ -	$ 172,212	$ (184,708)
	LDA	0.8024	68.61 (67.97,69.25)	$ 289,662	$ (15,532)	$ (226,461)	$ -	$ 47,670	$ (241,993)
	Logit	0.8136	70.97 (70.34,71.59)	$ 304,229	$ (17,497)	$ (211,894)	$ -	$ 74,838	$ (229,391)
	Average	**0.8258**	**74.92 (74.33,75.5)**	**$ 345,434**	**$ (18,221)**	**$ (170,689)**	**$ -**	**$ 156,524**	**$ (188,910)**
SMOTE	C5.0	0.8546	84 (83.49,84.5)	$ 465,004	$ (46,744)	$ (51,119)	$ -	$ 367,142	$ (97,863)
	CART	0.7300	77.9 (77.32,78.46)	$ 391,443	$ (51,751)	$ (124,680)	$ -	$ 215,012	$ (176,431)
	LDA	0.7569	73.16 (72.55,73.76)	$ 373,992	$ (41,461)	$ (142,131)	$ -	$ 190,400	$ (183,592)
	Logit	0.7866	75.93 (75.34,76.52)	$ 373,971	$ (32,778)	$ (142,152)	$ -	$ 199,041	$ (174,930)
	Average	**0.7820**	**77.75 (77.18,78.31)**	**$ 401,103**	**$ (43,183)**	**$ (115,021)**	**$ -**	**$ 242,899**	**$ (158,204)**
ROSE	C5.0	0.7299	35.06 (34.41,35.71)	$ 47,655	$ (445)	$ (468,468)	$ -	$ (421,258)	$ (468,913)
	CART	0.6543	51.15 (50.47,51.84)	$ 109,901	$ (4,644)	$ (406,223)	$ -	$ (300,966)	$ (410,867)
	LDA	0.7995	63.78 (63.12,64.44)	$ 212,286	$ (10,875)	$ (303,837)	$ -	$ (102,426)	$ (314,712)
	Logit	0.8042	66.77 (66.12,67.41)	$ 243,232	$ (13,491)	$ (272,891)	$ -	$ (43,150)	$ (286,382)
	Average	**0.7470**	**54.19 (53.53,54.85)**	**$ 153,268**	**$ (7,364)**	**$ (362,855)**	**$ -**	**$ (216,950)**	**$ (370,219)**
Overall Average		**0.7875**	**72 (71.41,72.59)**	**$ 342,922**	**$ (34,451)**	**$ (173,201)**	**$ -**	**$ 135,271**	**$ (207,651)**

However, without looking at the financial measures the retailer would have likely chosen the Up-C5.0 combination as their final production model to base their decision-support because it has the greatest test statistics (AUC = 0.9150/Accuracy = 85.48%). However, this model does not necessarily lead to the best expected profit or least misclassification costs.

The Raw-C5.0 combination revealed respectable test statistics (AUC = 0.7561/Accuracy = 83.6%) and led to an expected profit gain of $62,153 and reduced misclassification costs of $6,409 compared to using Up-C5.0.

Copyright © 2015, MatthewALanham.com

Conclusions and Future Research

Conclusions

The results of the study show the value that might be achieved using stochastic cost-benefit analysis for binary classification model assessment and selection. The motivation for this research is that the assortment decision is considered one of the most important decisions that a retailer will make and thus must consider the financial considerations that their stocking decisions could make. Using naïve approaches such as binary classification to identify SKU sellers from non-sellers could be a viable baseline modeling strategy when the number of SKUs to evaluate is in the hundreds of thousands. However, we posit that retailer's should also use more sophisticated approaches that incorporate substitution behavior.

Limitations

Our study is limited in that only one category of products is investigated in insolation, substitution behavior was disregarded, the evaluation of the business performance is retrospective over the same time frame as the predictions, and modeling build time is not incorporated. Ideally, all these concerns must be addressed by the retailer when generating timely and reliable decision-support.

Turning this work into a Journal-quality Paper

We want to extend this research by addressing these important financial aspects as well as other realistic concerns that a retailer must face so that complete assortment decision-support system can be created and tested.

Factors we are testing include incorporating:

(1) constructing a valid simulation design
(2) more sophisticated financial measures
(3) including substitution behavior
(4) evaluation over all categories that must be considered in a store assortment

Copyright © 2015, MatthewALanham.com

Acknowledgments

The authors would like to thank the retailer, decision-makers, and data scientists we have collaborated with for helping us to better understand the assortment planning process. This project has allowed us an opportunity to identify missing areas in the academic literature, but more importantly improve business practice by following the management science/analytics process.



 Copyright © 2015, MatthewALanham.com